B.P. DEMIDOVICH

ÉLÉMENTS DE CALCUL NUMÉRIQUE

ÉLÉMENTS DE CALCUL NUMÉRIQUE

PAR

B. DÉMIDOVITCH et I. MARON

ÉDITIONS MIR · MOSCOU

PRÉFACE

Les progrès impétueux des techniques nouvelles et l'application toujours plus poussée des mathématiques modernes à la recherche ont rendu infiniment plus rigoureux les impératifs auxquels doit satisfaire la formation mathématique des ingénieurs et des chercheurs se consacrant à des problèmes appliqués.

Les connaissances mathématiques nécessaires à un chercheur dans le domaine technique ne peuvent déjà plus se borner aux éléments traditionnels de l'analyse dite classique dont les branches essentielles se sont constituées en principe vers le début du XXe siècle. Un ingénieur travaillant dans un centre de recherches est actuellement censé de connaître de nombreuses branches des mathématiques modernes et, en premier lieu, de posséder à fond les méthodes et procédés de calcul numérique, la résolution de la plupart des problèmes imposant l'obtention d'un résultat numérique.

Les techniques actuelles offrent de nouveaux moyens puissants pour la réalisation effective des calculs. Dans de nombreux cas il devient donc possible de remplacer les positions approchées des problèmes appliqués par des formulations précises. Or cela implique la mise en œuvre de théories mathématiques spéciales (équations différentielles non linéaires, analyse fonctionnelle, méthodes variationnelles, etc.).

L'utilisation raisonnable des machines à calculer modernes est impensable sans la maîtrise des méthodes de calcul approché et numérique. C'est ce qui explique l'intérêt de plus en plus croissant porté en U.R.S.S. et dans d'autres pays aux méthodes d'analyse numérique.

Le présent ouvrage a surtout pour objet de donner un exposé dans une certaine mesure systématique et mis à jour des plus importants méthodes et procédés de calcul numérique sur la base d'un cours général des mathématiques supérieures. Il est conçu dans sa majeure partie comme un manuel de premier cycle d'étude de calcul numérique à l'usage des élèves des écoles techniques supérieures.

L'essor de l'économie nationale de l'U.R.S.S. rend nécessaire la formation en nombre toujours plus grand de spécialistes pour les centres de calcul. Les programmes de divers cours de perfectionnement des ingénieurs et de la formation des candidats à la licence réservent une grande place au calcul approché et numérique. C'est la raison pour laquelle nous avons inclus dans ce livre des renseignements supplémentaires sortant du cadre d'un cours d'école supérieure usuel. Cela ne rend pas pour autant plus difficile la lecture de l'ouvrage, le lecteur peut sans altérer la compréhension du texte choisir les exposés nécessaires et omettre ceux qu'il considère comme superflus. Pour rendre plus commode la manipulation du livre, les chapitres pouvant être omis en première lecture sont affectés d'un astérisque.

Les auteurs font largement appel aux principes du calcul matriciel. Les notions de vecteur, matrice, matrice inverse, valeur propre et vecteur propre, etc., sont usuelles. L'application des matrices présente plusieurs avantages, rendant plus facile la mise en évidence des principes de nombreux calculs. Dans ce sens la démonstration des théorèmes de convergence de divers processus numériques est particulièrement suggestive. En outre, les machines à calculer rapides actuelles réalisent sans peine les opérations matricielles principales.

Pour l'intelligence du texte le lecteur doit posséder un minimum de connaissances dans le domaine de l'algèbre linéaire et de la théorie des espaces vectoriels. L'assimilation de ce minimum devient plus facile grâce aux renseignements supplémentaires que les auteurs ont cru nécessaire d'inclure afin d'éviter des références à de nombreuses sources. Les chapitres respectifs sont indépendants du texte principal et le lecteur initié peut les omettre sans inconvénient.

Le présent ouvrage traite surtout des sujets suivants: opérations sur les nombres approchés; calcul des valeurs d'une fonction à l'aide de séries et de processus itératifs; résolution approchée et numérique des équations algébriques et transcendantes; méthodes numériques de l'algèbre linéaire; interpolation des fonctions; dérivation et intégration numérique des fonctions; méthode de Monte-Carlo.

Les auteurs s'attachent à donner des méthodes commodes pour évaluer les erreurs. Les théorèmes de convergence sont démontrés pour la plupart des processus, l'exposé rendant possible leur omission pour se borner seulement à l'aspect technique de la question. Dans certains cas, pour rendre le texte plus direct et moins complexe, les procédés de calcul sont présentés sous une forme schématisée.

Les méthodes principales sont poussées jusqu'à des applications numériques exposées sous la forme d'exemples et de résolutions détaillées. Pour une meilleure maîtrise des principes la majorité des exemples est traitée sous une forme simplifiée donnée à titre

d'illustration. Les références et la bibliographie supplémentaire sont placées en fin de chaque chapitre.

Le présent ouvrage est un exposé des méthodes choisies de calcul numérique et ne contient pas les textes relatifs aux formules empiriques, à l'approximation quadratique des fonctions, aux solutions approchées des équations différentielles, etc. Les auteurs se proposent d'écrire à ce sujet un ouvrage spécial.

Ce livre ne traite pas non plus de programmation et de techniques de résolution des problèmes mathématiques sur les machines à calculer.

Les auteurs tiennent à exprimer leur reconnaissance aux collaborateurs de la chaire des mathématiques supérieures de l'Académie d'Artillerie F. Dzerjinski, qui ont pris part à la discussion du manuscrit. Tout particulièrement nous remercions L. Lusternik, G. Tolstov, N. Bouslenko pour leurs remarques d'ordre général, E. Chouvalova qui a mis à notre disposition certaines données, D. Grobman pour des conseils pratiques précieux et A. Iouchkévitch qui a revu le chapitre XVII.

La reconnaissance des auteurs va également au professeur K. Smolitski, au chargé de cours S. Frolov et à R. Chostak dont les critiques ont permis d'améliorer le manuscrit.

Les Auteurs

INTRODUCTION

Généralités

La réalisation d'un grand nombre de calculs impose l'observation des règles bien simples élaborées par la pratique, qui facilitent le travail du calculateur et rendent rationnelle l'utilisation des machines et des moyens auxiliaires.

Le calculateur doit dresser en premier lieu un *schéma de calcul* qui indique exactement l'ordre des opérations et qui permet d'obtenir le résultat recherché par le moyen le plus simple et le plus rapide. Cela importe surtout dans le cas des calculs de même type, car alors un tel schéma, en rendant les calculs automatiques, permet de les exécuter à une vitesse et avec une fiabilité plus grandes, ce qui compense largement le temps nécessaire pour la composition du schéma. Par ailleurs, un schéma de calcul détaillé permet de confier le travail à des exécutants moins qualifiés.

Voici un exemple pour illustrer l'établissement d'un schéma. Supposons qu'il faille calculer les valeurs d'une fonction donnée analytiquement

$$y = f(x)$$

pour les valeurs données de l'argument $x = x_1, x_2, \ldots, x_n$. Si le nombre de ces valeurs est grand, il n'est pas raisonnable de calculer d'abord la valeur $f(x_1)$, puis la valeur $f(x_2)$, etc., réalisant chaque fois l'ensemble des opérations désignées par le symbole f. Il est beaucoup plus avantageux, après avoir décomposé la fonction f en opérations élémentaires

$$f(x) = f_m(\ldots(f_2(f_1(x)))\ldots),$$

de réaliser les calculs par des opérations de même type:

$$u_i = f_1(x_i) \qquad (i = 1, 2, \ldots, n);$$
$$v_i = f_2(u_i) \qquad (i = 1, 2, \ldots, n);$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y = f_m(w_i) \qquad (i = 1, 2, \ldots, n),$$

en reprenant toujours la même opération f_j ($j = 1, 2, \ldots, m$) pour toutes les valeurs considérées de l'argument. On peut alors utiliser largement les tables des fonctions correspondantes et les machines à calculer spécialisées. Les résultats des calculs doivent être portés sur des *cartes* ou *formulaires* spéciaux, feuilles de papier dûment réglées et portant les notations définies par le schéma de calcul retenu. A mesure que les résultats des calculs intermédiaires sont obtenus, on les inscrit sur ces cartes à l'endroit bien déterminé, ainsi que les résultats définitifs.

Les cartes sont conçues en général de façon que les résultats de chaque série des opérations de même type soient portés sur la même colonne ou ligne, la disposition des écritures des résultats intermédiaires devant être commode pour la réalisation des calculs ultérieurs.

Ainsi, pour composer le tableau des valeurs de la fonction

$$y = \frac{e^x + \cos x}{1 + x^2} + \sqrt{1 + \sin^2 x}, \tag{1}$$

on peut recommander le formulaire représenté sur le tableau 1.

Les calculs se font suivant les colonnes, le caractère des opérations de même type à réaliser étant suggéré par le formulaire luimême.

Tableau 1

Carte de calcul de la fonction (1)

x	x^2 (1)²	e^x	$\sin x$	$\cos x$	$e^x + \cos x$ (3) + (5)	$1 + x^2$ (1) + (2)	$\frac{e^x + \cos x}{1 + x^2}$ (6) : (7)	$\sin^2 x$ (4)²	$1 + \sin^2 x$ (1) + (9)	$\sqrt{1 + \sin^2 x}$ $\sqrt{(10)}$	y (8) + (11)
1	2	3	4	5	6	7	8	9	10	11	12

On inscrit d'abord sur la colonne (1) les valeurs données de l'argument x. Ensuite tous les nombres de la colonne (1) sont élevés au carré et portés sur la colonne (2). Puis pour chaque nombre de la colonne (1) on définit d'après les tables les valeurs successives de e^x, $\sin x$, $\cos x$, inscrites respectivement sur les colonnes (3), (4), (5).

Les colonnes suivantes indiquent les résultats des opérations intermédiaires. Par exemple, la colonne (6) donne la valeur de la somme $e^x + \cos x$ (schématiquement (3) + (5)), etc. Dans la dernière colonne figurent les valeurs de la fonction cherchée y. Lorsque

la forme de la carte est correcte; le calculateur ne recourt pas pendant le travail à la formule de calcul, en concentrant toute son attention pour remplir les colonnes.

Notons que le schéma de calcul et la forme de la carte sont intimement liés aux appareils utilisés et aux tableaux auxiliaires. Ainsi, dans certains cas, des résultats intermédiaires isolés ne sont pas portés sur la carte étant conservés par la mémoire de la machine. Parfois il est commode de considérer des ensembles standards des opérations comme une opération isolée. Par exemple, avec une règle à calcul la valeur numérique de l'expression du type

$$\frac{ab}{c}$$

peut être calculée sans fixer le résultat intermédiaire et il ne faut donc pas la décomposer en opérations élémentaires de multiplication et de division. D'une façon analogue, lorsqu'on travaille sur des calculateurs électriques, le calcul de la somme

$$\sum_{k=1}^{n} a_k b_k$$

est une opération unique. Dans de nombreux cas on a intérêt à transformer les expressions données pour les ramener à une forme singulière (par exemple, remplacer la division par la multiplication par une grandeur inverse ou décomposer l'expression en un produit commode pour le calcul des logarithmes, etc.).

Le deuxième élément auquel il faut prêter attention c'est la *vérification des calculs.* Sans cette vérification le calcul ne peut être considéré comme terminé. La vérification peut être *courante* ou *finale.* Les opérations supplémentaires de vérification courante permettent d'établir avec une certitude plus ou moins grande que les résultats intermédiaires obtenus sont valides. S'il n'en est pas ainsi, on reprend les calculs du pas correspondant. La vérification finale ne porte que sur le résultat définitif. Par exemple, si l'on cherche la racine d'une équation, la valeur obtenue peut être vérifiée par substitution. Le bon sens suggère que lorsque les calculs sont nombreux, il est trop risqué de remettre la vérification à la fin des opérations. C'est pourquoi il est plus avantageux de vérifier la validité des calculs pas à pas. Dans des cas importants les calculs sont vérifiés par exécution indépendante des opérations par deux calculateurs ou le calcul se fait par le même exécutant mais suivant deux méthodes différentes.

Le troisième élément important est l'*estimation de l'erreur.* Dans la majorité des cas les calculs se font avec des nombres approchés, le résultat obtenu étant aussi approché. C'est pourquoi même une méthode exacte de résolution des problèmes donne lieu à des *erreurs générées* (*erreurs d'opération*) et *des erreurs d'arrondi.* Si la méthode

elle-même est approchée, à ces deux erreurs vient s'ajouter l'*erreur de la méthode*. Dans des circonstances défavorables, l'erreur résultante peut être si grande que le résultat obtenu n'aura qu'une valeur illusoire. Les chapitres correspondants du présent ouvrage indiquent les méthodes d'évaluation des erreurs pour les calculs principaux.

Dans une carte de calcul il est utile de prévoir des colonnes pour les différences tabulées (cf. chapitre XIV, § 2) qu'on peut mettre à profit pour la vérification des calculs. Notamment, si la validité d'une partie du tableau des différences est compromise, il faut recalculer les éléments correspondants du tableau ou révéler la cause de la perturbation.

Il faut également veiller à ce que les inscriptions sur les cartes soient soignées et bien nettes. La pratique atteste qu'une écriture vague des chiffres conduit souvent à des fautes susceptibles de compromettre un calcul bien organisé. Les fautes d'écriture des nombres sont particulièrement graves si ces derniers comptent beaucoup de zéros. Les nombres de ce type doivent être notés sous une forme normale, en spécifiant la puissance entière de dix; par exemple

$$0{,}00000345 = 3{,}45 \cdot 10^{-6},$$

etc.

Dans ce livre, les auteurs traitent essentiellement des m é t h o d e s d e c a l c u l s. Dans plusieurs cas les exemples numériques sont simplifiés et les calculs intermédiaires souvent omis.

CHAPITRE PREMIER

NOMBRES APPROCHÉS

§ 1. Erreurs absolue et relative

Un *nombre approché* a est un nombre légèrement différent du nombre exact A et qui dans les calculs remplace ce dernier. Si l'on sait que $a < A$, a est dit valeur approchée du nombre A *par défaut*; si $a > A$, a est une valeur approchée *par excès*. Par exemple, pour $\sqrt{2}$ le nombre 1,41 est une valeur approchée par défaut alors que 1,42 l'est par excès car $1{,}41 < \sqrt{2} < 1{,}42$. Si a est une valeur approchée du nombre A, on note $a \approx A$.

Il est d'usage d'entendre par *erreur* Δa d'un nombre approché a la différence entre le nombre exact A correspondant et le nombre approché donné

$$\Delta a = A - a\ ^*.$$

Si $A > a$, l'erreur est positive: $\Delta a > 0$; mais si $A < a$, l'erreur est négative: $\Delta a < 0$. Pour obtenir le nombre exact A, il faut ajouter au nombre approché a l'erreur Δa

$$A = a + \Delta a.$$

Le nombre exact peut être ainsi considéré comme approché avec une erreur nulle.

Dans de nombreux cas le signe de l'erreur est inconnu. Il convient alors de recourir à l'*erreur absolue du nombre approché*

$$\Delta = |\,\Delta a\,|.$$

D é f i n i t i o n 1. On appelle *erreur absolue* Δ d'un nombre approché a la valeur absolue de la différence entre le nombre exact correspondant A et le nombre a

$$\Delta = |\,A - a\,|. \tag{1}$$

Il convient alors de distinguer deux cas:

1) le nombre A est connu; l'erreur absolue se détermine d'après la formule (1);

---

* On appelle quelquefois erreur la différence $a - A$.

2) le nombre A est inconnu, cas pratiquement le plus fréquent; par conséquent, nous ne pouvons pas déterminer l'erreur absolue Δ d'après la formule (1).

Dans ce cas, au lieu de l'erreur absolue théorique Δ inconnue, il est utile d'introduire sa limite supérieure dite *borne supérieure d'erreur absolue* *.

Définition 2. On désigne par le nom de *borne supérieure d'erreur absolue* d'un nombre approché tout nombre supérieur ou égal à l'erreur absolue de ce nombre.

Ainsi, si Δ_a est une borne d'erreur absolue d'un nombre approché a qui remplace le nombre exact A, on a

$$\Delta = |A - a| \leqslant \Delta_a. \tag{2}$$

On en déduit que le nombre exact A est encadré comme suit

$$a - \Delta_a \leqslant A \leqslant a + \Delta_a. \tag{3}$$

Par conséquent, $a - \Delta_a$ est une approche du nombre A par défaut et $a + \Delta_a$ l'est par excès.

Dans ce cas on fait appel à une écriture abrégée

$$A = a \pm \Delta_a.$$

Exemple 1. Trouver la borne d'erreur absolue du nombre $a = 3{,}14$ qui remplace le nombre π.

Solution. Puisqu'on a l'inégalité

$$3{,}14 < \pi < 3{,}15, \quad \text{il vient} \quad |a - \pi| < 0{,}01$$

et, par conséquent, on peut poser $\Delta_a = 0{,}01$.

Si l'on tient compte de ce que

$$3{,}14 < \pi < 3{,}142,$$

une meilleure estimation est $\Delta_a = 0{,}002$.

Constatons que la notion de borne d'erreur absolue énoncée dans ce qui précède est très large: *par borne d'erreur absolue d'un nombre approché a on entend un représentant quelconque de l'ensemble infini des nombres non négatifs Δ_a qui vérifient l'inégalité* (2). Il en résulte que tout nombre supérieur à une borne d'erreur absolue du nombre approché donné peut s'appeler également borne d'erreur absolue de ce nombre. En pratique il est commode de choisir pour Δ_a le nombre le plus petit possible qui vérifie l'inégalité (2).

L'écriture d'un nombre approché obtenu par mesure directe traduit en général sa borne d'erreur absolue. Par exemple, si la longueur d'un segment est $l = 214$ cm à 0,5 cm près, on écrit $l =$

---

* Sauf mention expresse, on entend ici par borne d'erreur la borne d'erreur supérieure (*note du trad.*).

$= 214$ cm $\pm$ 0,5 cm. Ici la borne d'erreur absolue $\Delta_l = 0{,}5$ cm, et la valeur exacte de la longueur l du segment est comprise entre les limites 213,5 cm $\leqslant l \leqslant$ 214,5 cm.

L'erreur absolue (ou la borne d'erreur absolue) ne suffit pas pour caractériser le degré de précision de la mesure ou du calcul. Ainsi, si en mesurant les longueurs de deux tiges on obtient $l_1 =$ $= 100{,}8$ cm $\pm$ 0,1 cm et $l_2 = 5{,}2$ cm $\pm$ 0,1 cm, bien que les bornes d'erreur absolue coïncident, la première mesure est meilleure que la deuxième. Pour la précision des données d'une mesure, le rôle essentiel revient à l'erreur absolue par unité de longueur qui s'appelle *erreur relative*.

D é f i n i t i o n 3. *L'erreur relative* δ d'un nombre approché a est le rapport de l'erreur absolue Δ de ce nombre et du module du nombre exact correspondant A $(A \neq 0)$

$$\delta = \frac{\Delta}{|A|}. \tag{4}$$

D'où $\Delta = |A|\delta$.

Introduisons, de même que pour l'erreur absolue, la notion de *borne d'erreur relative*.

D é f i n i t i o n 4. *La borne* (*supérieure*) *d'erreur relative* δ_a d'un nombre approché a donné est un nombre quelconque supérieur ou égal à l'erreur relative de ce nombre. Par définition :

$$\delta \leqslant \delta_a, \tag{5}$$

c'est-à-dire $\frac{\Delta}{|A|} \leqslant \delta_a$, d'où $\Delta \leqslant |A|\delta_a$.

Ainsi on peut prendre pour borne d'erreur absolue du nombre a

$$\Delta_a = |A|\delta_a. \tag{6}$$

Comme pratiquement $A \approx a$, au lieu de la formule (6) on utilise souvent la formule

$$\Delta_a = |a|\delta_a. \tag{6'}$$

Si l'on connaît une borne d'erreur relative δ_a, on en déduit un encadrement du nombre exact. Voici la notation conventionnelle qui traduit le fait que le nombre exact repose entre $a(1 - \delta_a)$ et $a(1 + \delta_a)$:

$$A = a(1 \pm \delta_a).$$

Soient a un nombre approché qui remplace le nombre exact A et Δ_a une borne d'erreur absolue du nombre a. Pour fixer les idées, posons que $A > 0$, $a > 0$ et $\Delta_a < a$. Alors

$$\delta = \frac{\Delta}{A} \leqslant \frac{\Delta_a}{a - \Delta_a}.$$

Par conséquent, on peut adopter comme borne d'erreur relative du nombre a le nombre

$$\delta_a = \frac{\Delta_a}{a - \Delta_a}.$$

D'une façon analogue on obtient $\Delta = A\delta \leqslant (a + \Delta)\, \delta_a$; d'où

$$\Delta_a = \frac{a\delta_a}{1 - \delta_a}.$$

Si, comme d'ordinaire, $\Delta_a \ll a$ et $\delta_a \ll 1$ (le symbole $\ll$ se lit « très inférieur »), on peut adopter d'une façon approchée

$$\delta_a \approx \frac{\Delta_a}{a}$$

et

$$\Delta_a \approx a\delta_a.$$

Exemple 2. Le poids de 1 dm³ d'eau à 0 °C est $p = 999,847$ gf $\pm$ 0,001 gf. Trouver la borne d'erreur relative du résultat de la pesée.

Solution. Il est clair que $\Delta_p = 0,001$ gf et $p = 999,846$ gf. Par conséquent,

$$\delta_p = \frac{0,001}{999,846} \approx 10^{-4}\ \%.$$

Exemple 3. En recherchant la constante de gaz de l'air on a obtenu $R = 29,25$. L'erreur relative de cette valeur étant 1‰, trouver les limites entre lesquelles est comprise R.

Solution. On a $\delta_R = 0,001$; donc $\Delta_R = R\delta_R \approx 0,03$. Par suite, $29,22 \leqslant R \leqslant 29,28$.

§ 2. Sources principales des erreurs

Les erreurs commises dans les problèmes mathématiques peuvent être en principe classées en cinq catégories.

1. Erreurs dues à la position même du problème. Il est rare que les formulations mathématiques présentent un modèle fidèle du phénomène réel; généralement, ces modèles sont plus ou moins idéalisés. Lors de l'étude de tel ou tel phénomène de la nature, dans les cas courants on est contraint, afin de simplifier le problème, d'admettre certaines conditions, ce qui donne lieu à plusieurs erreurs (*erreurs du problème*).

Il arrive parfois qu'il soit difficile ou même impossible de résoudre un problème énoncé en termes exacts. On le remplace alors par un problème approché dont les résultats se distinguent peu de ceux du problème donné. On fait apparaître ainsi une erreur qu'on peut appeler *erreur de la méthode*.

2. Erreurs associées en analyse mathématique aux processus infinis. Les fonctions qui figurent dans les formules mathématiques sont souvent données sous la forme de suites infinies ou de séries (par exemple, $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$). Plus même, de nombreuses équations mathématiques ne peuvent être résolues qu'en décrivant des processus infinis dont les limites constituent précisément les solutions cherchées. Un processus infini ne se terminant pas en général en un nombre fini de pas, on est obligé d'y mettre fin à un certain terme de la suite en le considérant comme une approximation de la solution cherchée. On comprend que le processus ainsi arrêté donne lieu à une erreur dite *erreur de troncature*.

3. Erreurs dues à la présence dans les formules mathématiques des paramètres numériques dont les valeurs ne peuvent être déterminées qu'approximativement. Telles sont, par exemple, toutes les constantes physiques. Cette erreur est dite par convention *initiale*.

4. Erreurs associées au système de numération. Même des nombres rationnels notés dans le système décimal ou dans un autre système positionnel peuvent comporter à droite de la virgule une infinité de chiffres (un nombre décimal périodique, par exemple). Il est évident que le calcul ne peut mettre en œuvre qu'un nombre fini de ces chiffres. Ainsi apparaît l'*erreur d'arrondi* ou *de chute*. En posant, par exemple, $\frac{1}{3} = 0{,}333$, on commet une erreur $\Delta \approx \approx 3 \cdot 10^{-4}$. Il arrive également qu'il faut arrondir des nombres décimaux à grand nombre de chiffres.

5. Erreur résultante d'une opération due aux erreurs des termes initiaux (*erreur propagée*). Il est clair qu'en effectuant des calculs sur des nombres approchés, les erreurs des données de départ sont propagées en quelque sorte sur le résultat des calculs. Sous ce rapport, les erreurs propagées sont i n s u r m o n t a b l e s.

Il va de soi que des erreurs de telle ou telle espèce n'interviennent pas nécessairement dans la résolution des problèmes concrets ou leur influence est négligeable. Mais généralement, pour une analyse complète il faut prendre en considération toutes les erreurs. Dans ce qui suit nous allons nous borner essentiellement au calcul des erreurs générées et des erreurs de la méthode.

§ 3. Notation décimale des nombres approchés. Chiffres significatifs. Nombre de chiffres exacts

On sait que tout nombre positif a peut être représenté sous la forme d'un nombre décimal de développement limité ou illimité

$$a = \alpha_m 10^m + \alpha_{m-1} 10^{m-1} + \alpha_{m-2} 10^{m-2} + \ldots + \alpha_{m-n+1} 10^{m-n+1} + \ldots, \quad (1)$$

où α_i sont les chiffres du nombre a ($\alpha_i = 0, 1, 2, \ldots, 9$), le chiffre de l'ordre le plus élevé $\alpha_m \neq 0$ et m est un entier (rang supérieur

du nombre a). Par exemple,

$$3141{,}59\ldots = 3\cdot 10^3 + 1\cdot 10^2 + 4\cdot 10^1 + 1\cdot 10^0 + 5\cdot 10^{-1} + 9\cdot 10^{-2} + \ldots$$

Le poids d'une unité varie en fonction de sa position dans le développement décimal (1) du nombre a. L'unité en première place vaut 10^m, en deuxième 10^{m-1}, en n-ième 10^{m-n+1}, etc.

Dans la pratique on n'utilise essentiellement que des nombres approchés qui sont des nombres décimaux limités

$$b = \beta_m 10^m + \beta_{m-1} 10^{m-1} + \ldots + \beta_{m-n+1} 10^{m-n+1} \quad (\beta_m \neq 0). \qquad (2)$$

Tous les chiffres conservés β_i ($i = m,\ m-1,\ \ldots,\ m-n+1$) s'appellent *chiffres significatifs* du nombre approché b. Il se peut bien que certains d'entre eux (à l'exception de β_m) soient nuls. La représentation positionnelle du nombre b dans le système de numération décimale impose quelquefois l'introduction des zéros au début ou à la fin du nombre. Par exemple,

$$b = 7\cdot 10^{-3} + 0\cdot 10^{-4} + 1\cdot 10^{-5} + 0\cdot 10^{-6} = \underline{0{,}00}7010,$$

ou

$$b = 2\cdot 10^9 + 0\cdot 10^8 + 0\cdot 10^7 + 3\cdot 10^6 + 0\cdot 10^5 = 2\,003\,0\underline{00\,000}.$$

Les zéros de ce type (soulignés dans nos exemples) ne sont pas considérés comme des chiffres significatifs.

Définition 1. On appelle *chiffre significatif* d'un nombre approché tout chiffre dans sa représentation décimale différent du zéro et un zéro s'il se trouve entre deux chiffres significatifs ou s'il constitue un chiffre conservé. Tous les autres zéros faisant partie du nombre approché et ne servant que pour désigner les rangs ne sont pas chiffres significatifs.

Par exemple, dans le nombre 0,002 080 les trois premiers zéros ne sont pas significatifs puisqu'ils ne servent qu'à indiquer les rangs des autres chiffres. Quant aux deux zéros qui suivent, ils sont significatifs, le premier étant placé entre les chiffres significatifs 2 et 8 et le second, comme l'indique la notation, traduit le fait que le nombre approché a conservé la décimale 10^{-6}. Si le dernier chiffre du nombre 0,002 080 n'est pas significatif, ce nombre se mettrait sous la forme 0,002 08. Dans cette optique, les nombres 0,002 080 et 0,002 08 ne sont pas équivalents, le premier comptant quatre chiffres significatifs et le second rien que trois.

Dans l'écriture de grands nombres les zéros à droite peuvent soit désigner les chiffres significatifs, soit définir le rang des chiffres restants. C'est pourquoi la notation usuelle des nombres peut donner lieu à des confusions. Par exemple, nous ne pouvons pas juger d'après la forme du nombre 689 000 combien de chiffres significatifs

compte-t-il. On peut affirmer seulement qu'il y en a au moins trois. Pour lever cette indétermination il faut expliciter l'ordre supérieur du nombre et l'écrire sous la forme $6{,}89 \cdot 10^6$ s'il compte trois chiffres significatifs, ou $6{,}8900 \cdot 10^6$ s'il en compte cinq, etc. En général, une écriture de ce type est commode pour les nombres qui comportent un grand nombre de zéros non significatifs, par exemple $0{,}000\,000\,120 = 1{,}20 \cdot 10^{-7}$, etc.

Introduisons la notion de *chiffres exacts d'un nombre décimal approché.*

Définition 2. On dit que les n premiers chiffres significatifs d'un nombre approché sont *exacts* si l'erreur absolue de ce nombre ne dépasse pas une demi-unité du rang du n-ième chiffre significatif en comptant de gauche à droite.

Ainsi, si l'on sait que pour un nombre approché a (1), qui remplace le nombre exact A,

$$\Delta = |A - a| \leqslant \frac{1}{2} \cdot 10^{m-n+1},$$

alors, par définition, les premiers n chiffres $\alpha_m, \alpha_{m-1}, \ldots, \alpha_{m-n+1}$ de ce nombre sont exacts.

Par exemple, pour le nombre exact $A = 35{,}97$ le nombre $a = 36{,}00$ est une approximation avec trois chiffres exacts du fait que $|A - a| = 0{,}03 < \frac{1}{2} \cdot 0{,}1$.

Constatons que tous les chiffres significatifs des tables mathématiques sont exacts. Ainsi les tables de logarithmes à cinq décimales assurent que l'erreur absolue de la mantisse ne dépasse pas $\frac{1}{2} \cdot 10^{-5}$, etc.

Le terme « n chiffres exacts » ne doit pas être pris à la lettre, c'est-à-dire au sens que les n premiers chiffres significatifs d'un nombre approché a donné, qui compte n chiffres exacts, coïncident avec les chiffres respectifs du nombre précis A. Par exemple, le nombre approché $a = 9{,}995$ qui remplace le nombre précis $A = 10$ compte trois chiffres exacts, tous ces chiffres étant différents. Toutefois, les cas sont nombreux où les chiffres exacts d'un nombre approché sont précisément les mêmes que les chiffres respectifs du nombre précis.

Remarque. Dans certains cas il est commode de dire que le nombre a est une approximation du nombre précis A à *n chiffres exacts* dans un sens lâche, en entendant par là que l'erreur absolue $\Delta = |A - a|$ ne dépasse pas une unité de rang du n-ième chiffre du nombre approché.

Par exemple, pour le nombre précis $A = 412{,}3567$, le nombre $a = 412{,}356$ est une approximation à six chiffres exacts dans un sens lâche du fait que $\Delta = 0{,}0007 < 1 \cdot 10^{-3}$.

Par la suite nous entendons, sauf mention du contraire, les chiffres exacts d'un nombre approché dans le sens de la définition 2 (c'est-à-dire *au sens strict*).

§ 4. Arrondissement des nombres

Considérons un nombre approché ou précis a écrit sous la forme décimale. Il arrive souvent qu'il faut l'*arrondir*, c'est-à-dire le remplacer par un nombre a_1 à plus petit nombre de chiffres significatifs. Le nombre a_1 est choisi de façon à minimiser l'*erreur d'arrondi* $| a_1 - a |$.

Règle d'arrondissement (supplémentaire). Pour arrondir un nombre jusqu'à n chiffres significatifs on rejette tous les chiffres à droite du n-ième chiffre significatif ou, s'il faut conserver les rangs, on les remplace par des zéros. Dans ces conditions:

1) si le premier des chiffres rejetés est inférieur à 5, les chiffres restent inchangés;

2) si le premier des chiffres rejetés est supérieur à 5, on ajoute une unité au dernier chiffre restant;

3) si le premier des chiffres rejetés est égal à 5 et si parmi les autres chiffres rejetés il y en a des non nuls, le dernier chiffre restant est augmenté de l'unité;

3a) mais si le premier des chiffres rejetés est égal à 5 alors que tous les autres chiffres rejetés sont des zéros, le dernier chiffre conservé reste sans changer s'il est pair ou on lui ajoute une unité s'il est impair (règle du chiffre pair).

Autrement dit, si en arrondissant un nombre on rejette moins d'une demi-unité de dernier rang conservé, les chiffres de tous les rangs conservés restent inchangés; mais si la partie rejetée du nombre est supérieure à une demi-unité du dernier rang conservé, on ajoute une unité au chiffre de ce rang. Dans des cas exceptionnels, lorsque la partie supprimée est égale exactement à une demi-unité du dernier rang conservé, pour que les erreurs d'arrondi se compensent, on fait appel à la règle du chiffre pair.

Il est évident qu'en appliquant la règle d'arrondissement l'erreur d'arrondi ne dépasse pas $\frac{1}{2}$ de l'unité du rang du dernier chiffre significatif conservé.

Exemple 1. En arrondissant le nombre

$$\pi = 3{,}1415926535 \ldots$$

jusqu'à cinq, quatre et trois chiffres significatifs, on obtient les nombres approchés 3,1416; 3,142; 3,14 avec des erreurs absolues inférieures à $\frac{1}{2}\cdot 10^{-4}$, $\frac{1}{2}\cdot 10^{-3}$ et $\frac{1}{2}\cdot 10^{-2}$.

Exemple 2. En arrondissant le nombre 1,2500 à deux chiffres significatifs, on obtient le nombre approché 1,2 avec une erreur absolue égale à $\frac{1}{2}\cdot 10^{-1} = 0{,}05$.

La précision du nombre approché dépend non pas du nombre de chiffres significatifs, mais du nombre de chiffres significatifs exacts [1], [2]. Lorsqu'un nombre approché comporte un nombre superflu de chiffres significatifs inexacts, on recourt à l'arrondissement. On se guide généralement sur la règle pratique suivante: *lors de l'exécution des calculs approchés, le nombre de chiffres significatifs des résultats intermédiaires ne doit pas dépasser de plus d'une ou de deux unités le nombre de chiffres exacts.* Le résultat définitif ne doit contenir plus d'un chiffre significatif excédentaire par rapport aux chiffres exacts. Si dans ce cas l'erreur absolue du résultat ne dépasse pas deux unités du dernier rang conservé, le chiffre excédentaire est dit *douteux*.

La règle énoncée permet, sans porter atteinte à la précision des calculs, d'éviter l'écriture des chiffres superflus et de réduire nettement la durée du calcul. La raison de la conservation des chiffres de réserve est que dans les cas courants, l'estimation des erreurs des résultats se fait pour les pires des variantes et l'erreur réelle peut être nettement inférieure à l'erreur théorique maximale. Ainsi, dans de nombreux cas les chiffres significatifs considérés comme inexacts sont en fait exacts.

Conformément à la précision de l'ensemble des calculs on est également amené à arrondir les nombres précis dont le nombre de chiffres significatifs est trop grand ou infini.

Notons que si le nombre précis A est arrondi d'après la règle supplémentaire à n chiffres significatifs, le nombre approché a ainsi obtenu compte n chiffres exacts (au sens strict).

Mais si un nombre approché a comptant n chiffres exacts est arrondi à n chiffres significatifs, l'approche a_1 nouvellement obtenue n'aura, en général, n chiffres exacts qu'au sens lâche. En effet, en vertu de l'inégalité

$$|A - a_1| \leqslant |A - a| + |a - a_1|$$

la borne d'erreur absolue du nombre a_1 se compose de l'erreur absolue du nombre a et de l'erreur d'arrondi.

§ 5. Relation entre l'erreur relative d'un nombre approché et le nombre de chiffres exacts

Démontrons le théorème qui met en liaison la valeur de l'erreur relative d'un nombre approché et le nombre de chiffres exacts de ce nombre [3], [4].

T h é o r è m e. *Si un nombre approché positif a compte n chiffres exacts au sens strict, son erreur relative* δ *ne dépasse pas le quotient de* $\left(\frac{1}{10}\right)^{n-1}$ *par le premier chiffre significatif de ce nombre, c'est-à-dire*

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1},$$

où α_m *est le premier chiffre significatif du nombre a.*

D é m o n s t r a t i o n. Soit

$$a = \alpha_m 10^m + \alpha_{m-1}10^{m-1} + \ldots + \alpha_{m-n+1}10^{m-n+1} + \ldots \qquad (\alpha_m \geqslant 1)$$

une valeur approchée du nombre précis A, qui compte n chiffres exacts. On a par définition:

$$\Delta = |A - a| \leqslant \frac{1}{2}\cdot 10^{m-n+1};$$

d'où l'on tire

$$A \geqslant a - \frac{1}{2}\cdot 10^{m-n+1}.$$

Cette dernière inégalité devient encore plus forte si on remplace le nombre a par un nombre inférieur $\alpha_m 10^m$

$$A \geqslant \alpha_m 10^m - \frac{1}{2}\cdot 10^{m-n+1} = \frac{1}{2}\cdot 10^m\left(2\alpha_m - \frac{1}{10^{n-1}}\right). \qquad (1)$$

Le deuxième membre de l'inégalité (1) devient minimal avec $n = 1$. Par suite

$$A \geqslant \frac{1}{2}\cdot 10^m(2\alpha_m - 1), \qquad (2)$$

ou, puisque

$$2\alpha_m - 1 = \alpha_m + (\alpha_m - 1) \geqslant \alpha_m,$$

on a

$$A \geqslant \frac{1}{2}\alpha_m 10^m.$$

Par conséquent,

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2}10^{m-n+1}}{\frac{1}{2}\alpha_m 10^m} = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Donc,

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \qquad (3)$$

Le théorème est démontré.

R e m a r q u e 1. En utilisant l'inégalité (2) on peut obtenir une estimation plus précise de l'erreur relative δ.

C o r o l l a i r e 1. On peut prendre comme borne d'erreur relative du nombre a

$$\delta_a = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}, \tag{4}$$

où α_m est le premier chiffre significatif du nombre a.

C o r o l l a i r e 2. Si le nombre a compte au moins deux chiffres exacts, c'est-à-dire si $n \geqslant 2$, il vérifie pratiquement la formule

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \tag{5}$$

En effet, avec $n \geqslant 2$ le terme $\frac{1}{10^{n-1}}$ de l'inégalité (1) peut être négligé. Il vient

$$A \geqslant \frac{1}{2}\cdot 10^m \cdot 2\alpha_m = \alpha_m 10^m,$$

ce qui entraîne

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2}\cdot 10^{m-n+1}}{\alpha_m 10^m} = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Par conséquent,

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

R e m a r q u e 2. Si le nombre approché a compte n chiffres exacts au sens lâche, les estimations (4) et (5) doivent être doublées.

E x e m p l e 1. Quelle est la borne d'erreur relative si au lieu du nombre π on prend le nombre $a = 3{,}14$?

S o l u t i o n. Dans le cas considéré $\alpha_m = 3$ et $n = 3$. Par conséquent,

$$\delta_a = \frac{1}{2\cdot 3}\left(\frac{1}{10}\right)^{3-1} = \frac{1}{600} = \frac{1}{6}\,\%.$$

E x e m p l e 2. Avec combien de chiffres faut-il calculer $\sqrt{20}$ pour que l'erreur ne dépasse pas 0,1 %?

S o l u t i o n. Le premier chiffre étant 4, $\alpha_m = 4$, de plus, $\delta = 0{,}001$. On a $\frac{1}{4\cdot 10^{n-1}} \leqslant 0{,}001$, d'où $10^{n-1} \geqslant 250$ et $n \geqslant 4$.

Le théorème énoncé permet d'obtenir l'erreur relative δ d'un nombre approché

$$a = \alpha_m \cdot 10^m + \alpha_{m-1} 10^{m-1} + \ldots \tag{6}$$

d'après le nombre de ses chiffres exacts.

Pour résoudre le problème inverse qui consiste à définir le nombre n de chiffres exacts du nombre (6) si l'on connaît son erreur relative δ, on recourt généralement à la formule approchée

$$\delta = \frac{\Delta}{a} \quad (a > 0),$$

où Δ est l'erreur absolue du nombre a. On en tire

$$\Delta = a\delta. \tag{7}$$

En tenant compte du rang supérieur du nombre Δ, on établit aisément le nombre de chiffres exacts du nombre approché donné a. En particulier, si

$$\delta \leqslant \frac{1}{10^n},$$

les formules (6) et (7) entraînent

$$\Delta \leqslant (\alpha_m + 1)\, 10^m \cdot 10^{-n} \leqslant 10^{m-n+1},$$

c'est-à-dire le nombre a compte au moins n décimales exactes au sens lâche. D'une façon analogue, si

$$\delta \leqslant \frac{1}{2 \cdot 10^n},$$

le nombre a compte n chiffres exacts au sens strict.

Exemple 3. La précision relative du nombre approché $a = 24\,253$ est 1 %. Combien de chiffres exacts compte-t-il?

Solution. On a

$$\Delta = 24\,253 \cdot 0{,}01 \approx 243 = 2{,}43 \cdot 10^2.$$

Par conséquent, seuls les deux premiers chiffres du nombre a sont exacts ($n = 2$); le chiffre des centaines est douteux. D'après la règle énoncée dans ce qui précède, il est préférable d'écrire le nombre a sous la forme $a = 2{,}4\underline{3} \cdot 10^4$.

Remarque. Le mode indiqué de détermination du nombre de chiffres exacts est approché. L'évaluation précise des chiffres exacts du nombre a doit être guidée par les inégalités

$$\delta \geqslant \frac{\Delta}{a + \Delta}$$

et

$$\Delta \leqslant \frac{a\delta}{1 - \delta} \quad (0 \leqslant \delta < 1).$$

§ 6. Tables des valeurs de la borne d'erreur relative en fonction du nombre de chiffres exacts et tables inverses

Si un nombre approché s'écrit avec les chiffres exacts indiqués, sa borne d'erreur relative se calcule sans peine. Le calcul de ce type est très fréquent, on a donc intérêt à rationaliser cette opération. Le tableau 2 [5] indique l'erreur relative en pour cent du nombre approché en fonction du nombre de chiffres exacts au sens lâche et des deux premiers chiffres significatifs du nombre en comptant de gauche à droite.

Soit, par exemple, le nombre approché 0,00354 à trois chiffres exacts. Etant donné qu'ici $n = 3$ et le nombre 35 est contenu dans l'intervalle entre 35, . . ., 39 sur le tableau 2, nous trouvons $\delta =$ $= 0{,}29$ %.

Tableau 2

Erreur relative (en %) des nombres à n chiffres exacts

Deux premiers chiffres significatifs	n			Deux premiers chiffres significatifs	n		
	2	3	4		2	3	4
10-11	10	1	0,1	35, ..., 39	2,9	0,29	0,029
12-13	8,3	0,83	0,083	40, ..., 44	2,5	0,25	0,025
14, ..., 16	7,1	0,71	0,071	45, ..., 49	2,2	0,22	0,022
17, ..., 19	5,9	0,59	0,059	50, ..., 59	2	0,2	0,02
20, ..., 22	5	0,5	0,05	60, ..., 69	1,7	0,17	0,017
23, ..., 25	4,3	0,43	0,043	70, ..., 79	1,4	0,14	0,014
26, ..., 29	3,8	0,38	0,038	80, ..., 89	1,2	0,12	0,012
30, ..., 34	3,3	0,33	0,033	90, ..., 99	1,1	0,11	0,011

Si l'on ne connaît que le premier chiffre du nombre, 4 par exemple, on prend évidemment le plus grand des nombres 2,5 et 2,2 qui correspondent aux variantes possibles 40, . . ., 44 et 45, . . ., 49 (avec $n = 2$). Si l'on ne connaît non plus le premier chiffre, on prend les nombres de la première ligne (10 % ; 1 % ; 0,1 %) comme les plus grands. Le tableau montre que trois chiffres exacts assurent une précision relative (au moins 1 %) suffisante pour la plupart des calculs techniques. Constatons que si le nombre approché compte deux, trois ou quatre chiffres exacts au sens strict, tous les nombres du tableau sont à diviser par deux.

Le tableau 3 [5] donne les bornes supérieures des erreurs relatives (en %) qui assurent à la valeur approchée donnée tel ou tel nombre de chiffres exacts au sens lâche en fonction de ses deux premiers chiffres.

Montrons sur un exemple comment il faut utiliser le tableau 3. Soit le nombre approché $a = 5{,}297$ d'erreur relative $\delta = 0{,}5$ %.

Tableau 3

Nombre de chiffres exacts d'un nombre approché en fonction de la borne d'erreur relative (en %)

Deux premiers chiffres significatifs	n			Deux premiers chiffres significatifs	n		
	2	3	4		2	3	4
10-11	4,2	0,42	0,042	35, ..., 39	1,2	0,12	0,012
12-13	3,6	0,36	0,036	40, ..., 44	1,1	0,11	0,011
14, ..., 16	2,9	0,29	0,029	45, ..., 49	1	0,1	0,01
17, ..., 19	2,5	0,25	0,025	50, ..., 54	0,9	0,09	0,009
20, ..., 22	2,2	0,22	0,022	55, ..., 59	0,8	0,08	0,008
23, ..., 25	1,9	0,19	0,019	60, ..., 69	0,7	0,07	0,007
26, ..., 29	1,7	0,17	0,017	70, ..., 79	0,6	0,06	0,006
30, ..., 34	1,4	0,14	0,014	80, ..., 99	0,5	0,05	0,005

Les deux premiers chiffres significatifs sont 5 et 2; le nombre formé par ces chiffres est compris entre 50 et 54; de plus, les erreurs relatives associées à ces derniers en fonction du nombre de chiffres exacts sont 0,9 %; 0,09 %; 0,009 %, etc. Etant donné que $\delta = 0,5\ \% < 0,9\ \%$ et que l'erreur relative d'un nombre ne dépend pas des rangs des chiffres de ce nombre, le nombre $a = 5,297$ compte deux chiffres exacts au sens lâche.

Exemples. 1. En posant $\pi = 3,142$; $\sqrt{7} = 2,65$; $e = 2,718$; $\lg 5 = 0,699$; $\sin 1° = 0,0174$, trouvons dans le tableau 2 les erreurs relatives correspondantes:

$$\delta = 0,033\ \%;\quad \delta = 0,19\ \%;\quad \delta = 0,019\ \%;$$
$$\delta = 0,17\ \%;\quad \delta = 0,59\ \%.$$

2. Le module de Young $E = 2212\ldots$ tf/cm² est calculé à 2 % près d'après la flèche d'une tige d'acier. Quel est le nombre de chiffres exacts de la valeur obtenue? Le tableau 3 donne $n = 2$. Par conséquent, $E = 22\cdot 10^2$ tf/cm².

3. La constante de gaz $R = 31,5\ldots$ du mélange carburant d'un moteur à gaz est calculée avec une erreur relative $\delta = 1$ %. Trouver le nombre de chiffres exacts. D'après le tableau 3, $n = 2$. Donc, $R = 32$.

§ 7. Erreur d'une somme

Théorème 1. *L'erreur absolue d'une somme algébrique de plusieurs nombres approchés ne dépasse pas la somme des erreurs absolues de ces nombres.*

Démonstration. Soient $x_1, x_2, \ldots, x_n$ des nombres approchés donnés. Considérons leur somme algébrique

$$u = \pm x_1 \pm x_2 \pm \ldots \pm x_n.$$

Il est clair que

$$\Delta u = \pm \Delta x_1 \pm \Delta x_2 \pm \ldots \pm \Delta x_n$$

et, par conséquent,

$$| \Delta u | \leqslant | \Delta x_1 | + | \Delta x_2 | + \ldots + | \Delta x_n |. \qquad (1)$$

C o r o l l a i r e. On peut adopter que la borne d'erreur absolue d'une somme algébrique est la somme des bornes d'erreurs absolues des termes de la somme

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}. \qquad (2)$$

La formule (2) entraîne que la borne d'erreur absolue d'une somme ne peut être inférieure à la borne d'erreur absolue du terme le moins précis (en erreur absolue), c'est-à-dire du terme dont l'erreur absolue est maximale. Par conséquent, quel que soit le degré de précision de la détermination des autres termes ils ne peuvent pas apporter une contribution essentielle à la précision de la somme. Il s'ensuit qu'il n'y a aucune raison de conserver les chiffres superflus dans les termes plus précis. On en déduit une règle pratique d'addition des nombres approchés.

R è g l e. Pour additionner les nombres de précision absolue différente il faut:

1) souligner les nombres dont le développement décimal s'arrête avant celui des autres et les laisser sans changer;

2) arrondir les autres nombres conformément aux nombres soulignés en conservant un ou deux chiffres de réserve;

3) additionner les nombres en tenant compte de tous les chiffres conservés;

4) arrondir d'un chiffre le résultat obtenu.

Si l'on applique la règle supplémentaire pour arrondir les termes de la somme

$$u = x_1 + x_2 + \ldots + x_n$$

au rang m, l'erreur d'arrondi de la somme ne dépasse pas dans le pire des cas la valeur

$$\Delta_{\text{ar}} \leqslant n \cdot \frac{1}{2} \cdot 10^m. \qquad (3)$$

Un calcul plus précis de l'erreur d'arrondi d'une somme s'obtient si l'on tient compte des signes des erreurs d'arrondi des termes additionnés.

E x e m p l e. Chercher la somme des nombres approchés: 0,348; 0,1834; 345,4; 235,2; 11,75; 9,27; 0,0849; 0,0214; 0,000354 dont tous les chiffres sont significatifs (au sens lâche).

S o l u t i o n. Ecrivons les nombres les moins exacts 345,4 et 235,2, dont l'erreur absolue peut atteindre 0,1. En arrondissant

les autres nombres à 0,01 près, on a

$$\begin{array}{r} 345,4 \\ 235,2 \\ 11,75 \\ 9,27 \\ 0,35 \\ 0,18 \\ 0,08 \\ 0,02 \\ 0,00 \\ \hline 602,25 \end{array}$$

En arrondissant le résultat à 0,1 près d'après la règle du chiffre pair, la valeur approchée obtenue de la somme est 602,2.

L'erreur totale Δ du résultat se compose de trois termes dont

1) la somme des bornes d'erreur des données de départ

$$\Delta_1 = 10^{-3} + 10^{-4} + 10^{-1} + 10^{-1} + 10^{-2} + 10^{-2} + 10^{-4} + \\ + 10^{-4} + 10^{-6} = 0,221\,301 < 0,222;$$

2) la valeur absolue de la somme des erreurs d'arrondi (compte tenu de leurs signes) des termes

$$\Delta_2 = | -0,002 + 0,0034 + 0,0049 + 0,0014 + 0,000354 | = \\ = 0,008054 < 0,009;$$

3) l'erreur de l'arrondissement final

$$\Delta_3 = 0,050.$$

Par conséquent,

$$\Delta = \Delta_1 + \Delta_2 + \Delta_3 \leqslant 0,222 + 0,009 + 0,050 = 0,281 < 0,3;$$

et la somme cherchée est donc $602,2 \pm 0,3$.

Théorème 2. *Si les termes d'une somme sont du même signe, sa borne d'erreur relative ne dépasse pas la borne d'erreur relative maximale de ses termes.*

Démonstration. Soit $u = x_1 + x_2 + \ldots + x_n$, où, pour fixer les idées, $x_i > 0$ $(i = 1, 2, \ldots, n)$.

Désignons par A_i $(A_i > 0;\ i = 1, 2, \ldots, n)$ les valeurs exactes des termes x_i et par $A = A_1 + A_2 + \ldots + A_n$ la valeur exacte de la somme u. Alors on peut prendre comme borne d'erreur relative de la somme

$$\delta_u = \frac{\Delta_u}{A} = \frac{\Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}}{A_1 + A_2 + \ldots + A_n}. \qquad (4)$$

Puisque

$$\delta_{x_i} = \frac{\Delta_{x_i}}{A_i} \quad (i = 1, 2, \ldots, n),$$

on a

$$\Delta_{x_i} = A_i \delta_{x_i}. \tag{4'}$$

En portant cette expression dans la formule (4), on obtient:

$$\delta_u = \frac{A_1\delta_{x_1} + A_2\delta_{x_2} + \ldots + A_n\delta_{x_n}}{A_1 + A_2 + \ldots + A_n}.$$

Soit $\bar{\delta}$ la plus grande des erreurs relatives δ_{x_i}, c'est-à-dire $\bar{\delta}_{x_i} \leqslant \bar{\delta}$. Il vient

$$\delta_u \leqslant \frac{\bar{\delta}\,(A_1 + A_2 + \ldots + A_n)}{A_1 + A_2 + \ldots + A_n} = \bar{\delta}.$$

Par conséquent, $\delta_u \leqslant \bar{\delta}$, soit

$$\delta_u \leqslant \max(\delta_{x}, \delta_{x_2}, \ldots, \delta_{x_n}).$$

§ 8. Erreur d'une différence

Considérons la différence de deux nombres approchés $u = x_1 - x_2$.

D'après la formule (2) du § 7, la borne d'erreur absolue Δ_u de la différence

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2},$$

c'est-à-dire *la borne d'erreur absolue d'une différence est égale à la somme des bornes d'erreurs absolues de ses termes.*

On en tire la borne d'erreur relative

$$\delta_u = \frac{\Delta_{x_1} + \Delta_{x_2}}{A}, \tag{1}$$

où A est la valeur exacte de la valeur absolue de la différence des nombres x_1 et x_2.

R e m a r q u e s u r l'a l t é r a t i o n d e l a p r é c i s i o n d a n s l e c a s d e s o u s t r a c t i o n d e s n o m b r e s v o i s i n s. Si les nombres approchés x_1 et x_2 sont assez proches l'un de l'autre et si leurs erreurs absolues sont petites, le nombre A est petit. La formule (1) entraîne que dans ce cas la borne d'erreur relative peut être très grande alors que les erreurs relatives des termes de la différence restent faibles, c'est-à-dire on est en présence d'une *perte de précision.*

Calculons, par exemple, la différence de deux nombres $x_1 = 47{,}132$ et $x_2 = 47{,}111$ dont chacun compte cinq chiffres significatifs exacts. En retranchant on obtient $u = 47{,}132 - 47{,}111 = 0{,}021$.

La différence u ne comporte ainsi que deux chiffres significatifs dont le dernier est douteux, la borne d'erreur absolue de la différence étant

$$\Delta_u = 0{,}0005 + 0{,}0005 = 0{,}001.$$

Les bornes d'erreurs relatives du nombre à soustraire, du plus grand nombre et de la différence sont respectivement

$$\delta_{x_1} = \frac{0{,}0005}{47{,}132} \approx 0{,}00001\,;$$

$$\delta_{x_2} = \frac{0{,}0005}{47{,}111} \approx 0{,}00001\,;$$

$$\delta_u = \frac{0{,}001}{0{,}021} \approx 0{,}05.$$

La borne d'erreur relative de la différence est ici 5 000 fois environ plus grande que les bornes d'erreurs relatives des données initiales.

Pour cette raison dans les calculs approchés il est utile de transformer les expressions qui conduisent à la soustraction des nombres voisins.

E x e m p l e. Trouver la différence

$$u = \sqrt{2{,}01} - \sqrt{2} \tag{2}$$

avec trois chiffres exacts.

S o l u t i o n. Etant donné que

$$\sqrt{2{,}01} = 1{,}4177\,4469\ \ldots$$

et

$$\sqrt{2} = 1{,}4142\,1356\ \ldots,$$

le résultat cherché est

$$u = 0{,}00353 = 3{,}53 \cdot 10^{-3}.$$

Le même résultat s'obtient si l'expression (2) se met sous la forme

$$u = \frac{0{,}01}{\sqrt{2{,}01} + \sqrt{2}}$$

et si l'on limite les racines $\sqrt{2{,}01}$ et $\sqrt{2}$ aux trois chiffres exacts. En effet,

$$u = \frac{0{,}01}{1{,}42 + 1{,}41} = \frac{0{,}01}{2{,}83} = 10^{-2} \cdot 3{,}53 \cdot 10^{-1} = 3{,}53 \cdot 10^{-3}.$$

Les considérations précédentes permettent d'énoncer la règle pratique suivante: dans le calcul approché il convient d'éviter au

possible la soustraction de deux nombres approchés à peu pres égaux; mais si une telle soustraction s'impose, les termes de la différence doivent être pris avec un nombre suffisant de chiffres exacts de réserve (si cela est possible). Par exemple, si l'on sait qu'en retranchant x_2 de x_1 on fait disparaître les premiers m chiffres significatifs alors que le résultat à obtenir doit compter n chiffres significatifs exacts, x_1 et x_2 doivent être pris avec $m + n$ chiffres significatifs exacts.

§ 9. Erreur d'un produit

Théorème. *L'erreur relative d'un produit de plusieurs nombres approchés différents du zéro ne dépasse pas la somme des erreurs relatives de ces nombres.*

Démonstration. Soit $u = x_1x_2 \ldots x_n$.

Supposons pour simplifier que les nombres approchés $x_1, x_2, \ldots, x_n$ soient positifs; on a

$$\ln u = \ln x_1 + \ln x_2 + \ldots + \ln x_n.$$

D'où, en utilisant la formule approchée $\Delta \ln x \approx d \ln x = \frac{\Delta x}{x}$, on a

$$\frac{\Delta u}{u} = \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2} + \ldots + \frac{\Delta x_n}{x_n}.$$

L'évaluation de cette dernière expression en valeur absolue donne

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\Delta x_2}{x_2}\right| + \ldots + \left|\frac{\Delta x_n}{x_n}\right|.$$

Si A_i ($i = 1, 2, \ldots, n$) désignent les valeurs exactes des facteurs x_i et si $|\Delta x_i|$, comme il en est dans les cas courants, sont petits par rapport à x_i, on peut poser approximativement:

$$\left|\frac{\Delta x_i}{x_i}\right| \approx \left|\frac{\Delta x_i}{A_i}\right| = \delta_i$$

et

$$\left|\frac{\Delta u}{u}\right| = \delta,$$

où les δ_i sont les erreurs relatives des facteurs x_i ($i = 1, 2, \ldots, n$) et δ est l'erreur relative du produit.

Par conséquent,

$$\delta \leqslant \delta_1 + \delta_2 + \ldots + \delta_n. \qquad (1)$$

La formule (1) reste évidemment valide si les signes des facteurs x_i ($i = 1, 2, \ldots, n$) sont différents.

Corollaire. La borne d'erreur relative du produit est égale à la somme des bornes d'erreurs relatives des facteurs, c'est-

à-dire

$$\delta_u = \delta_{x_1} + \delta_{x_2} + \ldots + \delta_{x_n}. \quad (2)$$

Si tous les facteurs du produit u, sauf un, sont très précis, la formule (2) entraîne que la borne d'erreur relative du produit coïncide pratiquement avec la borne d'erreur relative du facteur le moins précis. Dans le cas particulier où seul le facteur x_1 est approché, on a simplement

$$\delta_u = \delta_{x_1}.$$

Si l'on connaît la borne d'erreur relative δ_u du produit u, on peut définir sa borne d'erreur absolue Δ_u d'après la formule

$$\Delta_u = |u| \delta_u.$$

E x e m p l e 1. Déterminer le produit u des nombres approchés $x_1 = 12{,}2$ et $x_2 = 73{,}56$ et le nombre de ses chiffres exacts, si tous les chiffres écrits des facteurs sont exacts.

S o l u t i o n. On a $\Delta_{x1} = 0{,}05$ et $\Delta_{x2} = 0{,}005$. Il en résulte

$$\delta_u = \frac{0{,}05}{12{,}2} + \frac{0{,}005}{73{,}56} = 0{,}0042.$$

Le produit de ces nombres étant $u = 897{,}432$, $\Delta_u = u\delta_u = 897 \cdot 0{,}004 = 3{,}6$ (approximativement).

u ne compte donc que deux chiffres exacts et le résultat doit s'écrire:

$$u = 897 \pm 4.$$

Signalons le cas particulier

$$u = kx$$

où k est un facteur exact différent du zéro. On a:

$$\delta_u = \delta_x$$

et

$$\Delta_u = |k| \Delta_x,$$

c'est-à-dire *lorsqu'on multiplie un nombre approché par un nombre exact k, la borne d'erreur relative ne change pas, alors que la borne d'erreur absolue devient $|k|$ fois plus grande.*

E x e m p l e 2. La borne d'erreur angulaire du pointage d'une fusée est $\varepsilon = 1'$. Quel est, sur une distance de 2 000 km, l'écart possible Δ_u de la fusée par rapport au but si une correction n'intervient pas?

S o l u t i o n. Ici

$$\Delta_u = \frac{\pi}{180 \cdot 60} \cdot 2\,000 \text{ km} \approx 580 \text{ m}.$$

Il est évident que l'erreur relative du produit ne peut pas être inférieure à l'erreur relative du facteur le moins précis. C'est pourquoi, de même que dans le cas de la somme, aucune raison n'est de conserver des chiffres exacts excédentaires des facteurs plus précis.

Il est utile de se guider sur la règle suivante: pour trouver le produit de plusieurs nombres approchés de nombres différents de chiffres significatifs exacts, il suffit:

1) de les arrondir de façon que chacun d'eux compte un ou deux chiffres significatifs de plus que le nombre de chiffres exacts du facteur le moins précis;

2) de conserver après multiplication autant de chiffres significatifs qu'il y a de chiffres exacts dans le facteur le moins précis (ou retenir un chiffre de réserve en plus).

E x e m p l e 3. Chercher le produit des nombres approchés $x_1 = 2{,}5$ et $x_2 = 72{,}397$ si les chiffres écrits sont exacts.

S o l u t i o n. En appliquant la règle, on a après l'arrondissement: $x_1 = 2{,}5$; $x_2 = 72{,}4$, ce qui donne $x_1x_2 = 2{,}5 \cdot 72{,}4 = 181 \approx$ $\approx 1{,}8 \cdot 10^2$.

§ 10. Nombre de chiffres exacts d'un produit

Soit un produit de n facteurs ($n \leqslant 10$) $u = x_1x_2 \ldots x_n$ dont chacun compte au moins m ($m > 1$) chiffres exacts. Soit, ensuite, $\alpha_1, \alpha_2, \ldots, \alpha_n$ les premiers chiffres significatifs du développement décimal des facteurs:

$$x_i = \alpha_i 10^{p_i} + \beta_i 10^{p_i - 1} + \ldots \quad (i = 1, 2, \ldots, n).$$

La formule (5) du § 5 amène alors

$$\delta_{x_i} = \frac{1}{2\alpha_i}\left(\frac{1}{10}\right)^{m-1} \quad (i = 1, 2, \ldots, n)$$

et, par conséquent.

$$\delta_u = \frac{1}{2}\left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n}\right)\left(\frac{1}{10}\right)^{m-1}. \tag{1}$$

Etant donné que $\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n} \leqslant 10$, on a $\delta_u \leqslant \frac{1}{2}\left(\frac{1}{10}\right)^{m-2}$.

Par conséquent, même dans le pire des cas, le produit u compte $m - 2$ chiffres exacts.

R è g l e. Si tous les facteurs comptent m décimales exactes et si leur nombre est inférieur ou égal à 10, le nombre de chiffres exacts (au sens lâche) d'un produit est d'une ou de deux unités inférieur à m.

Donc, si un produit doit contenir m chiffres exacts, les facteurs doivent être pris avec un ou deux chiffres de réserve.

Si la précision des facteurs est différente, il faut entendre par m le nombre de chiffres exacts du facteur le moins précis. Ainsi, *le nombre de chiffres exacts du produit d'un petit nombre de facteurs (de l'ordre de dix) peut être d'une ou de deux unités inférieur au nombre de chiffres exacts du facteur le moins précis.*

Exemple 1. Déterminer l'erreur relative et le nombre de chiffres exacts du produit $u = 93{,}87 \cdot 9{,}236$.

Solution. La formule (1) donne

$$\delta_u = \frac{1}{2}\left(\frac{1}{9}+\frac{1}{9}\right)\frac{1}{10^3} = \frac{1}{9}\cdot 10^{-3} < \frac{1}{2}\cdot 10^{-3}.$$

Par conséquent, le produit u compte au moins trois chiffres exacts (cf. § 5).

Exemple 2. Trouver l'erreur relative et le nombre de chiffres exacts du produit $u = 17{,}63 \cdot 14{,}285$.

Solution.

$$\delta_u = \frac{1}{2}\left(\frac{1}{1}+\frac{1}{1}\right)\frac{1}{10^3} = 1\cdot 10^{-3}.$$

Par conséquent, le produit compte au moins trois chiffres exacts (au sens lâche).

§ 11. Erreur d'un quotient

Si $u = \frac{x}{y}$, on a $\ln u = \ln x - \ln y$ et

$$\frac{\Delta u}{u} = \frac{\Delta x}{x} - \frac{\Delta y}{y}.$$

Il en résulte

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta y}{y}\right|.$$

Cette dernière formule implique que le théorème du § 9 s'étend à un quotient.

Théorème. *L'erreur relative d'un quotient ne dépasse pas la somme des erreurs relatives du dividende et du diviseur.*

Corollaire. Si $u = \frac{x}{y}$, on a $\delta_u = \delta_x + \delta_y$.

Exemple. Chercher le nombre de chiffres exacts du quotient $u = 25{,}7 : 3{,}6$ si tous les chiffres écrits du dividende et du diviseur sont exacts.

S o l u t i o n. On a

$$\delta_u = \frac{0,05}{25,7} + \frac{0,05}{3,6} = 0,002 + 0,014 = 0.016.$$

Comme $u = 7,14$, $\Delta_u = 0,016 \cdot 7,14 = 0,11$. Pour cette raison le quotient u comporte deux chiffres exacts au sens lâche, c'est-à-dire $u = 7,1$ ou, plus précisément,

$$u = 7,14 \pm 0,11.$$

§ 12. Nombre de chiffres exacts d'un quotient

Soient le dividende x et le diviseur y comptant chacun au moins m chiffres exacts. Si α et β sont leurs premiers chiffres significatifs, on peut adopter que la borne d'erreur relative du quotient u est le nombre

$$\delta_u = \frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)\left(\frac{1}{10}\right)^{m-1}.$$

On en tire la règle suivante: 1) si $\alpha \geqslant 2$ et $\beta \geqslant 2$, le quotient u compte au moins $m - 1$ chiffres exacts; 2) si $\alpha = 1$ ou $\beta = 1$, le quotient compte au moins $m - 2$ chiffres exacts.

§ 13. Erreur relative d'une puissance

Soit $u = x^m$ (m étant un nombre naturel), alors $\ln u = m \ln x$, donc

$$\left|\frac{\Delta u}{u}\right| = m \left|\frac{\Delta x}{x}\right|$$

et on tombe sur

$$\delta_u = m\delta_x, \tag{1}$$

c'est-à-dire *la borne d'erreur relative de la m-ième puissance d'un nombre est m fois plus grande que la borne d'erreur relative du nombre lui-même.*

§ 14. Erreur relative d'une racine

Soit maintenant $u = \sqrt[m]{x}$, alors $u^m = x$. Il vient

$$\delta_u = \frac{1}{m}\,\delta_x, \tag{1}$$

c'est-à-dire *la borne d'erreur relative de la m-ième racine est m fois plus petite que la borne d'erreur relative du radicande.*

E x e m p l e. Avec quelle erreur relative et avec combien de chiffres exacts peut-on déterminer la mesure du côté a d'un carré dont la surface $s = 12,34$ (à 0,01 près).

Solution. On a $a=\sqrt{s}=3{,}5128\ldots$ Puisque

$$\delta_s=\frac{0{,}01}{12{,}33}\approx 0{,}0008,$$

on a $\delta_a=\frac{1}{2}\delta_s=0{,}0004$. Donc

$$\Delta_a = 3{,}5128\cdot 0{,}0004 = 1{,}4\cdot 10^{-3}.$$

On en tire que le nombre a compte à peu près quatre chiffres exacts (au sens lâche) et, par conséquent, $a = 3{,}513$.

§ 15. Calculs sans estimation précise des erreurs

Dans les paragraphes précédents nous avons exposé les moyens permettant d'évaluer la borne d'erreur absolue d'une opération. Nous y avons supposé que les erreurs absolues des composantes se renforcent réciproquement, circonstance qui se produit rarement en pratique.

Dans le cas d'un très grand nombre de calculs, lorsqu'on ne tient pas compte de l'erreur de chaque résultat isolé, il est recommandé d'appliquer les règles suivantes d'établissement du nombre de chiffres [6].

1. Dans l'addition et la soustraction des nombres approchés, le rang inférieur conservé du résultat doit être égal au plus fort des rangs des derniers chiffres significatifs exacts des données initiales.

2. Dans la multiplication et la division des nombres approchés, il faut conserver dans le résultat autant de chiffres significatifs qu'il y en a dans la donnée approchée au nombre inférieur de chiffres significatifs exacts.

3. En élevant un nombre approché au carré ou au cube le résultat doit conserver autant de chiffres significatifs que compte de chiffres significatifs exacts la base de la puissance.

4. Lors de l'extraction d'une racine carrée ou cubique d'un nombre approché, il faut prendre le résultat avec autant de chiffres significatifs qu'il y a de chiffres exacts dans le radicande.

5. Tous les résultats intermédiaires doivent compter un chiffre en plus de ceux recommandés par les règles précédentes. Ce chiffre « de réserve » est à rejeter dans le résultat final.

6. Dans le cas du calcul à l'aide des logarithmes, il est recommandé d'utiliser une table des logarithmes au nombre de décimales d'une unité supérieur par rapport au plus petit nombre de chiffres significatifs exacts du nombre approché. Le dernier chiffre significatif du résultat est à rejeter.

7. Si les données peuvent être prises avec une précision arbitraire, pour obtenir le résultat avec k chiffres exacts, les données

de départ doivent être prises avec un nombre de chiffres tel qu'il assure dans le résultat final d'après les règles précédentes $k+1$ chiffres exacts.

Si certaines données comportent des rangs inférieurs excédentaires (addition et soustraction) ou plus de chiffres significatifs que d'autres données (multiplication, division, élévation à la puissance, extraction de la racine), il faut les arrondir au préalable en conservant un chiffre de réserve.

§ 16. Formule générale de l'erreur

Voici le problème principal de la théorie des erreurs: définir l'erreur de la fonction donnée de plusieurs grandeurs dont on connaît les erreurs.

Soit la fonction dérivable donnée

$$u = f(x_1, x_2, \ldots, x_n)$$

et soient $|\Delta x_i|$ $(i = 1, 2, \ldots, n)$ les erreurs absolues des arguments de la fonction. L'erreur absolue de la fonction est alors

$$|\Delta u| = |f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots$$
$$\ldots, x_n + \Delta x_n) - f(x_1, x_2, \ldots, x_n)|.$$

En pratique les $|\Delta x_i|$ sont généralement de petites grandeurs dont les produits, carrés et puissances supérieures peuvent être négligés. On peut donc poser:

$$|\Delta u| \approx |df(x_1, x_2, \ldots, x_n)| = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i\right| \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| |\Delta x_i|$$

si grad $f(x_1, x_2, \ldots, x_n) \neq 0$.

Ainsi

$$|\Delta u| \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| |\Delta x_i|. \quad (1)$$

On en tire, en désignant par Δ_{x_i} $(i = 1, 2, \ldots, n)$ les bornes d'erreurs absolues des arguments x_i et par Δ_u la borne d'erreur de la fonction u, pour des Δx_i petits:

$$\Delta_u = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \Delta_{x_i}. \quad (2)$$

Après avoir divisé par u les deux membres de l'inégalité (1) on obtient l'estimation de l'erreur relative de la fonction u

$$\delta \leqslant \sum_{i=1}^{n} \left|\frac{\frac{\partial f}{\partial x_i}}{u}\right| |\Delta x_i| = \sum_{i=1}^{n} \left|\frac{\partial}{\partial x_i} \ln f(x_1, \ldots, x_n)\right| |\Delta x_i|. \quad (3)$$

Par conséquent, on peut prendre comme borne d'erreur relative de la fonction u

$$\delta_u = \sum_{i=1}^{n} \left| \frac{\partial}{\partial x_i} \ln u \right| \Delta_{x_i}. \tag{4}$$

Exemple 1. Chercher les bornes d'erreurs absolue et relative du volume d'une sphère $V = \frac{1}{6} \pi d^3$, si le diamètre $d = 3{,}7$ cm $\pm$ $\pm 0{,}05$ cm et $\pi \approx 3{,}14$.

Solution. En considérant π et d comme des grandeurs variables, calculons les dérivées partielles

$$\frac{\partial V}{\partial \pi} = \frac{1}{6} d^3 = 8{,}44\,;$$

$$\frac{\partial V}{\partial d} = \frac{1}{2} \pi d^2 = 21{,}5.$$

En vertu de la formule (2), la borne d'erreur absolue du volume

$$\Delta_V = \left| \frac{\partial V}{\partial \pi} \right| |\Delta\pi| + \left| \frac{\partial V}{\partial d} \right| |\Delta d| = 8{,}44 \cdot 0{,}0016 + 21{,}5 \cdot 0{,}05 =$$

$$= 0{,}013 + 1{,}075 = 1{,}088 \text{ cm}^3 \approx 1{,}1 \text{ cm}^3.$$

Par suite

$$V = \frac{1}{6} \pi d^3 \approx 27{,}4 \text{ cm}^3 \pm 1{,}1 \text{ cm}^3. \tag{5}$$

Il en résulte la borne d'erreur relative du volume

$$\delta_V = \frac{1{,}088 \text{ cm}^3}{27{,}4 \text{ cm}^3} = 0{,}0397 \approx 4\ \%.$$

Exemple 2. Pour définir le module de Young E d'après la flèche d'une tige de section rectangulaire on emploie la formule

$$E = \frac{1}{4} \cdot \frac{l^3 p}{a^3 b s},$$

où l est la longueur de la tige, a et b les mesures de sa section transversale, s la flèche et p la charge.

Calculer la borne d'erreur relative de la détermination du module de Young E si $p = 20$ kgf; $\delta_p = 0{,}1\ \%$; $a = 3$ mm; $\delta_a = 1\ \%$; $b = 44$ mm; $\delta_b = 1\ \%$; $l = 50$ cm; $\delta_l = 1\ \%$; $s = 2{,}5$ cm; $\delta_s = 1\ \%$.

Solution. $\ln E = 3 \ln l + \ln p - 3 \ln a - \ln b - \ln s - - \ln 4$.

En remplaçant les accroissements par les différentielles, on a

$$\frac{\Delta E}{E}=3\frac{\Delta l}{l}+\frac{\Delta p}{p}-3\frac{\Delta a}{a}-\frac{\Delta b}{b}-\frac{\Delta s}{s}.$$

Donc,

$$\delta_E = 3\delta_l + \delta_p + 3\delta_a + \delta_b + \delta_s = 3\cdot 0{,}01 + 0{,}001 + \\ + 3\cdot 0{,}01 + 0{,}01 + 0{,}01 = 0{,}081.$$

Ainsi la borne d'erreur relative est 0,081, c'est-à-dire elle constitue 8 % environ de la grandeur mesurée.

Les calculs numériques nous conduisent à

$$E= 2{,}10 \pm 0{,}17)\ 10^6\ \frac{\text{kgf}}{\text{cm}^2}.$$

§ 17. Problème inverse de la théorie des erreurs

L'intérêt du problème inverse est également grand pour la pratique: ce problème consiste à chercher les erreurs absolues des arguments d'une fonction telles que l'erreur absolue de cette fonction ne dépasse pas la valeur imposée.

Ce problème est mathématiquement indéterminé puisque la même borne d erreur donnée Δ_u de la fonction $u = f(x_1, x_2, \ldots, x_n)$ peut s'obtenir à partir de plusieurs combinaisons de bornes d'erreurs absolues Δ_{x_i} de ses arguments.

La solution la plus simple du problème inverse est donnée par ce qu'on appelle le *principe d'égalité des effets.* D'après ce principe on suppose que la contribution de toutes les différentielles partielles

$$\frac{\partial f}{\partial x_i}\Delta_{x_i} \quad (i=1,\ 2,\ \ldots,\ n)$$

est la même dans la formation de l'erreur absolue Δ_u de la fonction $u = f(x_1, x_2, \ldots, x_n)$.

Soit la valeur de la borne d'erreur absolue Δ_u. La formule (2) du § 16 amène alors

$$\Delta_u=\sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}. \qquad (1)$$

En supposant que tous les termes soient égaux entre eux, on a

$$\left|\frac{\partial u}{\partial x_1}\right|\Delta_{x_1}=\left|\frac{\partial u}{\partial x_2}\right|\Delta_{x_2}=\ldots=\left|\frac{\partial u}{\partial x_n}\right|\Delta_{x_n}=\frac{\Delta_u}{n}.$$

Il en résulte

$$\Delta_{x_i}=\frac{\Delta_u}{n\left|\frac{\partial u}{\partial x_i}\right|} \quad (i=1,\ 2,\ \ldots,\ n). \qquad (2)$$

E x e m p l e 1. Le rayon de la base d'un cylindre est $R \approx 2$ m; la hauteur du cylindre $H \approx 3$ m. Quelles sont les erreurs absolues de la détermination de R et de H pour que son volume V soit calculé à 0,1 m³ près?

S o l u t i o n. On a $V = \pi R^2 H$ et $\Delta_V = 0{,}1$ m³.

En posant $R = 2$ m; $H = 3$ m; $\pi = 3{,}14$; on obtient approximativement:

$$\frac{\partial V}{\partial \pi} = R^2 H = 12;$$

$$\frac{\partial V}{\partial R} = 2\pi RH = 37{,}7;$$

$$\frac{\partial V}{\partial H} = \pi R^2 = 12{,}6.$$

Puisque $n = 3$, on en tire (d'après la formule (2))

$$\Delta_\pi = \frac{0{,}1}{3 \cdot 12} < 0{,}003;$$

$$\Delta_R = \frac{0{,}1}{3 \cdot 37{,}7} < 0{,}001;$$

$$\Delta_H = \frac{0{,}1}{3 \cdot 12{,}6} < 0{,}003.$$

E x e m p l e 2. Chercher la valeur de la fonction

$$u = 6x^2 (\lg x - \sin 2y)$$

avec deux décimales exactes (après la virgule), les valeurs approchées de x et de y étant respectivement égales à 15,2 et 57°. Trouver l'erreur absolue admissible de ces grandeurs.

S o l u t i o n. Ici

$$u = 6x^2 (\lg x - \sin 2y) = 6\,(15{,}2)^2 (\lg 15{,}2 - \sin 114^\circ) = 371{,}9;$$

$$\frac{\partial u}{\partial x} = 12x (\lg x - \sin 2y \quad . \; 6xM = 88{,}54,$$

où $M = \lg e = 0{,}43429$;

$$\frac{\partial u}{\partial y} = -12x^2 \cos 2y = +1127{,}7.$$

Pour que le résultat soit exact avec deux décimales, il faut que l'égalité $\Delta_u = 0{,}005$ soit vérifiée. Le principe d'égalité des effets entraîne alors

$$\Delta_x = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial x}\right|} = \frac{0{,}005}{2 \cdot 88{,}54} = 0{,}000028;$$

$$\Delta_y = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial y}\right|} = \frac{0{,}005}{2 \cdot 1127{,}7} = 0\ 0000022 \text{ rd} = 0'',45.$$

Il est fréquent qu'en résolvant un problème inverse à l'aide du principe d'égalité des effets les bornes d'erreurs absolues des variables indépendantes isolées définies d'après la formule (2) sont si petites que lors de la mesure de ces grandeurs il est pratiquement impossible d'obtenir la précision imposée. Dans ces cas il convient de renoncer au principe d'égalité des effets et en diminuant raisonnablement les erreurs d'une partie des variables les augmenter pour les variables de l'autre partie.

Exemple 3. Quels doivent être la précision de la mesure du rayon d'un cercle $R = 30{,}5$ cm et le nombre de chiffres de π pour obtenir la surface du cercle à 0,1 % près?

Solution. On a $s = \pi R^2$ et $\ln s = \ln \pi + 2 \ln R$. Il vient

$$\frac{\Delta_s}{s} = \frac{\Delta_\pi}{\pi} + \frac{2\Delta_R}{R} = 0{,}001.$$

D'après le principe d'égalité des effets il faut poser

$$\frac{\Delta_\pi}{\pi} = 0{,}0005; \quad \frac{2\Delta_R}{R} = 0{,}0005.$$

Il en résulte $\Delta_\pi \leqslant 0{,}0016$ et $\Delta_R \leqslant 0{,}00025R = 0{,}0076$ cm. Ainsi il faudrait prendre $\pi = 3{,}14$ et mesurer R à un millième de centimètre près. Cette précision est évidemment difficile à réaliser en pratique. C'est pourquoi il est plus avantageux de procéder de la façon suivante: prendre $\pi = 3{,}142$; d'où $\frac{\Delta_\pi}{\pi} = 0{,}00013$; alors $\frac{2\Delta_R}{R} = 0{,}001 - 0{,}00013 = 0{,}00087$ et $\Delta_R \leqslant 0{,}013$ cm. Cette précision s'obtient sans peine.

On admet parfois que la borne d'erreur absolue de tous les arguments x_i $(i = 1, 2, \ldots, n)$ est la même. Alors en posant

$$\Delta_{x_1} = \Delta_{x_2} = \ldots = \Delta_{x_n}$$

la formule (1) amène

$$\Delta_{x_i} = \frac{\Delta_u}{\sum\limits_{i=1}^{n} \left| \frac{\partial u}{\partial x_i} \right|} \qquad (i = 1, 2, \ldots, n).$$

Enfin on peut supposer que la précision de la mesure de tous les arguments x_i $(i = 1, 2, \ldots, n)$ soit la même, c'est-à-dire que les bornes d'erreurs relatives δ_{x_i} $(i = 1, 2, \ldots, n)$ des arguments soient égales entre elles:

$$\delta_{x_1} = \delta_{x_2} = \ldots = \delta_{x_n}.$$

On a donc

$$\frac{\Delta_{x_1}}{|x_1|} = \frac{\Delta_{x_2}}{|x_2|} = \ldots = \frac{\Delta_{x_n}}{|x_n|} = k,$$

où k est la valeur commune des rapports.

Par conséquent,

$$\Delta_{x_i} = k\,|x_i| \qquad (i = 1, 2, \ldots, n).$$

En portant ces valeurs dans la formule (1), on a

$$\Delta_u = k \sum_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|$$

et

$$k = \frac{\Delta_u}{\sum_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|}.$$

Finalement on obtient:

$$\Delta_{x_i} = \frac{|x_i|\,\Delta_u}{\sum_{j=1}^{n} \left| x_j \frac{\partial u}{\partial x_j} \right|} \qquad (i = 1, 2, \ldots, n)$$

On peut également utiliser d'autres variantes.

D'une façon analogue on résout le deuxième problème inverse de la théorie des erreurs lorsqu'on connaît la borne d'erreur relative de la fonction et qu'il faut chercher les bornes d'erreurs relative et absolue des arguments.

Quelquefois la formulation même du problème contient des conditions interdisant l'application du principe d'égalité des effets.

E x e m p l e 4. Les côtés d'un rectangle sont $a \approx 5$ m et $b \approx$ ≈ 200 m. Quelle doit être la borne d'erreur absolue admissible de la mesure, la même pour les deux côtés, pour que la surface S du rectangle soit définie avec une borne d'erreur absolue $\Delta_S = 1$ m^2?

S o l u t i o n. Etant donné que

$$S = ab,$$

on a

$$\Delta S \approx b\Delta a + a\Delta b$$

et

$$\Delta_S = b\Delta_a + a\Delta_b.$$

Par condition

$$\Delta_a = \Delta_b,$$

donc

$$\Delta_a = \frac{\Delta_S}{a+b} = \frac{1}{205} \approx 0{,}005 \text{ m} = 5 \text{ mm}.$$

§ 18. Précision de la détermination de l'argument d'une fonction tabulée

Il arrive souvent en pratique que l'argument doive être déterminé d'après la valeur tabulée de la fonction. Ainsi il est très fréquent qu'on se trouve devant la nécessité de chercher un nombre d'après son logarithme tabulé ou un angle d'après la valeur tabulée d'une quelconque de ses fonctions trigonométriques, etc. Une erreur de la fonction entraîne évidemment une erreur dans la détermination de l'argument.

Soit la table à une entrée de la fonction $y = f(x)$.

Si la fonction $f(x)$ est dérivable, on a pour les valeurs de $|\Delta x|$ suffisamment petites:

$$|\Delta y| = |f'(x)|\,|\Delta x|.$$

On en tire

$$|\Delta x| = \frac{|\Delta y|}{|f'(x)|}, \tag{1}$$

ou

$$\Delta_x = \frac{1}{|y'|}\Delta_y.$$

Appliquons la formule (1) à certaines fonctions tabulées les plus employées.

A. Fonctions logarithmiques

Soit $y = \ln x$, alors $y' = \frac{1}{x}$.

Il en résulte

$$\Delta_x = x\Delta_y. \tag{2}$$

Mais si $y = \lg x$, alors $y' = \frac{M}{x}$, où $M = 0{,}43429$;

$$\Delta_x = \frac{1}{M}x\Delta_y = 2{,}30x\Delta_y. \tag{2'}$$

On en déduit notamment que $\delta_x = 2{,}30\Delta_y$, c'est-à-dire la borne d'erreur relative d'un nombre dans la table des logarithmes décimaux est égale environ à 2,5 bornes d'erreur absolue du logarithme de ce nombre.

B. Fonctions trigonométriques

1. Si $y = \sin x \left(0 < x < \frac{\pi}{2}\right)$, alors $y' = \cos x$ et, par conséquent,

$$\Delta_x = \Delta_y \sec x \text{ rd}. \tag{3}$$

2. Pour la fonction

$$y = \operatorname{tg} x \left(0 < x < \frac{\pi}{2}\right)$$

on a

$$y' = \sec^2 x$$

et

$$\Delta_x = \Delta_y \cos^2 x \text{ rd}. \qquad (4)$$

3. Si $y = \lg(\sin x)\left(0 < x < \frac{\pi}{2}\right)$,

$$y' = M \operatorname{cotg} x \quad \text{et} \quad \Delta_x = 2{,}30 \operatorname{tg} x \Delta_y \text{ rd}. \qquad (5)$$

4. Posons $y = \lg(\operatorname{tg} x)\left(0 < x < \frac{\pi}{2}\right)$; il vient

$$y' = \frac{2M}{\sin 2x} \quad \text{et} \quad \Delta_x = 1{,}15 \sin 2x \Delta_y \text{ rd}. \qquad (6)$$

Puisqu'il est clair que $\frac{\sin 2x}{2} < \operatorname{tg} x$ pour $0 < x < \frac{\pi}{2}$, les formules (5) et (6) entraînent que, d'après la table des logarithmes des tangentes, l'angle x est établi avec une meilleure précision que d'après la table des logarithmes des sinus.

C. Fonction exponentielle

Si $y = e^x$, $y' = e^x$ et

$$\Delta_x = \frac{\Delta_y}{e^x} \qquad (7)$$

ou

$$\Delta_x = \frac{\Delta_y}{y}.$$

Exemple 1. Avec quelle précision peut-on déterminer le nombre $x \approx 5000$ si l'on utilise la table des logarithmes décimaux à quatre décimales?

Solution. La formule (2′) amène

$$\Delta_x = 2{,}30 \cdot 5000 \cdot \frac{1}{2} \cdot 10^{-4} \approx 0{,}6,$$

c'est-à-dire le nombre x compte environ quatre chiffres exacts.

Exemple 2. Trouver l'erreur de la définition de l'angle $x \approx 60°$:

a) d après la table des logarithmes des sinus à cinq décimales;
b) d'après la table des logarithmes des tangentes à cinq décimales.

Solution. Pour le premier cas on a d'après la formule (5):

$$\Delta_x = 2\ 30 \cdot \sqrt{3}\ \frac{1}{2} \cdot 10^{-5} \text{ rd} = 0{,}00002 \text{ rd} \approx 4''.$$

Dans le deuxième cas la formule (6) conduit à

$$\Delta_x = 1,15 \cdot \sqrt{3} \cdot \frac{1}{2} \cdot 10^{-5} \text{ rd} \approx 0,000005 \text{ rd} \approx 1'',$$

c'est-à-dire l'erreur est quatre fois plus petite.

§ 19. Méthode d'encadrement

Dans les cas courants l'erreur d'une fonction (§ 16, formule (2)) est évaluée approximativement parce que l'on néglige les produits des erreurs. Dans certains cas il faut connaître les bornes exactes de la valeur cherchée de la fonction si l'on connaît les limites de la variation de ses arguments. Pour l'obtenir, le plus simple est de faire appel à la méthode d'encadrement.

Soit

$$u = f(x_1, x_2, \ldots, x_n)$$

une fonction continûment dérivable, monotone par rapport à tout argument x_i $(i = 1, 2, \ldots, n)$. Pour l'obtenir, il suffit de supposer que les dérivées $\frac{\partial f}{\partial x_i}$ $(i = 1, 2, \ldots, n)$ conservent leur signe dans le domaine considéré ω de variation des arguments. Supposons que

$$\underline{x}_i < x_i < \overline{x}_i \quad (i = 1, 2, \ldots, n), \tag{1}$$

et que le parallélépipède (1) soit contenu complètement dans le domaine ω.

Posons que $\breve{x}_i = \underline{x}_i$, $\hat{x}_i = \overline{x}_i$ si la fonction f est croissante en x_i et $\breve{x}_i = \overline{x}_i$, $\hat{x}_i = \underline{x}_i$ si la fonction f est décroissante en x_i.

Il devient alors clair que

$$\underline{u} < u < \overline{u}, \tag{2}$$

où

$$\underline{u} = f(\breve{x}_1, \breve{x}_2, \ldots, \breve{x}_n)$$

et

$$\overline{u} = f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n).$$

Constatons que les variables $\breve{x}_i$ $(i = 1, 2, \ldots, n)$ et le résultat des opérations f sur ces variables ne peuvent être arrondis que dans le sens de décroissance de la grandeur $\underline{u}$, alors que les variables $\hat{x}_i$ $(i = 1, 2, \ldots, n)$ et le résultat des opérations f sur ces variables ne peuvent être arrondis que dans le sens de croissance de la grandeur $\overline{u}$. Si l'on remplit ces conditions, l'inégalité (2) est alors strictement observée. Dans le cas particulier d'une fonction f monotone croissante par rapport à tout argument x_i $(i = 1, 2, \ldots, n)$, on a sim-

plement

$$f(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n) < u < f(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n). \tag{3}$$

E x e m p l e. Un cylindre d'aluminium de diamètre de base $d = 2$ cm $\pm$ 0,01 cm et de hauteur $h = 11$ cm $\pm$ 0,02 cm pèse $p = 93{,}4$ gf $\pm$ 0,001 gf. Trouver le poids spécifique γ de l'aluminium et évaluer sa borne d'erreur absolue.

S o l u t i o n. Le volume du cylindre

$$v = \frac{\pi d^2}{4} h;$$

d'où

$$\gamma = \frac{p}{v} = \frac{4p}{\pi d^2 h}. \tag{4}$$

La formule (4) entraîne que dans le domaine $p > 0$, $d > 0$, $h > 0$ la fonction γ est croissante par rapport à l'argument p et décroissante par rapport aux arguments d et h. Par condition:

$$1{,}99 \text{ cm} \leqslant d \leqslant 2{,}01 \text{ cm};$$
$$10{,}98 \text{ cm} \leqslant h \leqslant 11{,}02 \text{ cm};$$
$$93{,}399 \text{ gf} \leqslant p \leqslant 93{,}401 \text{ gf}.$$

Par ailleurs

$$3{,}14159 < \pi < 3{,}1416.$$

C'est pourquoi

$$\underline{\gamma} = \frac{4 \cdot 93{,}399}{3{,}1416 \cdot 2{,}01^2 \cdot 11{,}02} = 2{,}671 \frac{\text{gf}}{\text{cm}^3}$$

(p a r d é f a u t) et

$$\overline{\gamma} = \frac{4 \cdot 93{,}401}{3{,}14159 \cdot 1{,}99^2 \cdot 10{,}98} = 2{,}735 \frac{\text{gf}}{\text{cm}^3}$$

(p a r e x c è s). En prenant la moyenne arithmétique, on a:

$$\gamma = 2{,}703 \frac{\text{gf}}{\text{cm}^3} \pm 0{,}027 \frac{\text{gf}}{\text{cm}^3} \tag{5}$$

et, après l'arrondissement,

$$\gamma = 2{,}70 \frac{\text{gf}}{\text{cm}^3} \pm 0{,}03 \frac{\text{gf}}{\text{cm}^3}.$$

A titre de comparaison voici une approximation de l'erreur. En utilisant les valeurs moyennes des arguments, on a

$$\gamma = \frac{4 \cdot 93{,}4}{3{,}1416 \cdot 2^2 \cdot 11} = 2{,}703 \frac{\text{gf}}{\text{cm}^3}.$$

En cherchant le logarithme des membres de la formule (4) on obtient:

$$\ln \gamma = \ln 4 + \ln p - \ln \pi - 2 \ln d - \ln h;$$

on en tire en prenant la différentielle totale:

$$\frac{\Delta\gamma}{\gamma}=\frac{\Delta p}{p}-\frac{\Delta\pi}{\pi}-\frac{2\Delta d}{d}-\frac{\Delta h}{h}.$$

Par conséquent,

$$\delta_\gamma=\delta_p+\delta_\pi+2\delta_d+\delta_h=\frac{0,001}{93,4}+\frac{0,00001}{3,1416}+\frac{2\cdot 0,01}{2}+\frac{0,02}{11}=$$

$$=1,07\cdot 10^{-5}+3,18\cdot 10^{-6}+10^{-2}+1,82\cdot 10^{-3}=1,183\cdot 10^{-2}.$$

Ensuite on obtient:

$$\Delta_\gamma=\delta_\gamma\cdot\gamma=1,183\cdot 10^{-2}\cdot 2,703=3,2\cdot 10^{-2}\,\frac{\mathrm{gf}}{\mathrm{cm}^3}.$$

Ainsi on a approximativement:

$$\gamma=2,703\,\frac{\mathrm{gf}}{\mathrm{cm}^3}\pm 0,032\,\frac{\mathrm{gf}}{\mathrm{cm}^3},$$

ce qui coïncide assez bien avec une estimation précise (5).

§ 20*. Notion de l'estimation probabiliste d'une erreur

Soit la somme de n termes

$$u=x_1+x_2+\ldots+x_n.$$

On sait que la borne d'erreur absolue est alors égale à

$$\Delta_u=\Delta_{x_1}+\Delta_{x_2}+\ldots+\Delta_{x_n}. \tag{1}$$

Et pour le cas de bornes d'erreurs absolues égales

$$\Delta_{x_1}=\Delta_{x_2}=\ldots=\Delta_{x_n}=\Delta,$$

on a:

$$\Delta_u=n\Delta. \tag{1'}$$

La formule (1) donne la valeur possible maximale de l'erreur absolue de la somme. Cette borne d'erreur n'est atteinte que si les erreurs de tous les termes: 1) sont les plus grandes possibles et 2) ont les mêmes signes. Dans le cas d'un grand nombre de termes ce concours de circonstances défavorable est peu probable. En règle générale, les erreurs réelles des termes isolés ont des signes différents et, par conséquent, elles se compensent partiellement. C'est pourquoi, en plus de la borne d'erreur théorique de la somme Δ_u, on introduit la *borne d'erreur pratique* Δ_u^* vraie dans une certaine mesure.

Bornons-nous à l'examen d'un cas bien simple. Supposons que les erreurs absolues Δx_i $(i=1, 2, \ldots, n)$ des termes de la somme (1) soient indépendants et vérifient la loi normale avec la même mesure de précision. Posons que les erreurs absolues des termes ne

dépassent pas le nombre Δ avec une probabilité supérieure au nombre γ, c'est-à-dire

$$P(|\Delta x_i| \leqslant \Delta) > \gamma.$$

Sous cette réserve on montre en calcul des probabilités que l'erreur absolue de la somme u vérifie avec la même mesure d'authenticité l'inégalité $|\Delta u| \leqslant \Delta\sqrt{n}$, où n est le nombre de termes.

On peut donc adopter que la borne d'erreur absolue d'une somme est donnée par le nombre

$$\Delta_u^* = \Delta\sqrt{n}. \qquad (2)$$

Par exemple, en additionnant 100 nombres avec une erreur absolue 0,1, on obtient la borne d'erreur théorique de la somme $\Delta_u = 0{,}1 \cdot 100 = 10$. Or, en fait on peut s'attendre à ce que cette erreur ne dépasse pas la quantité $0{,}1 \cdot 10 = 1$.

Considérons, notamment, la moyenne arithmétique de n nombres

$$\xi = \frac{1}{n}(x_1 + x_2 + \ldots + x_n).$$

D'après la théorie stricte, la borne d'erreur absolue

$$\Delta_\xi = \frac{1}{n} \cdot n\Delta = \Delta;$$

alors qu'on peut affirmer avec une grande authenticité que pratiquement

$$\Delta_\xi^* = \frac{\Delta\sqrt{n}}{n} = \frac{\Delta}{\sqrt{n}},$$

c'est-à-dire qu'il est pratiquement vrai *que la précision de la moyenne arithmétique des nombres approchés est plus grande que celle des nombres donnés et de plus*

$$\Delta_\xi^* \to 0 \quad \text{lorsque} \quad n \to \infty.$$

D'une façon analogue on peut montrer pour le cas d'un produit de n facteurs de même borne d'erreur relative δ que la borne d'erreur relative du produit est définie pratiquement par la formule

$$\delta_u^* = \delta\sqrt{n} \qquad (3)$$

BIBLIOGRAPHIE

1. *A. Krylov.* Conférences sur les calculs approchés. 2e éd. Académie des Sciences de l'U.R.S.S., Léningrad, 1933, chapitre I.
2. *D. Ventsel, E. Ventsel.* Eléments de la théorie des calculs approchés. Editions de l'Académie militaire de l'Air N. Joukovski, Moscou, 1949, chapitre I.
3. *J. Scarborough.* Numerical Mathematical Analysis. John Hopkins, 1950.
4. *I. Bésikovitch.* Calculs approchés. Gostekhizdat, 1949, chapitres I et II.
5. *G. Fikhtengoltz.* Mathématiques pour ingénieurs. GTTI, 1933, première partie, chapitre I.
6. *V. Bradis.* Calcul mental et écrit. Moyens auxiliaires de calcul. Encyclopédie des mathématiques élémentaires. Livre I. Outchpedguiz, 1951.

CHAPITRE II

GÉNÉRALITÉS SUR LA THÉORIE DES FRACTIONS CONTINUES

§ 1. Définition d'une fraction continue

L'expression du type

$$a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cfrac{b_3}{a_3+\ldots}}}=\left[a_0;\frac{b_1}{a_1},\frac{b_2}{a_2},\frac{b_3}{a_3},\ldots\right] \qquad (1)$$

s'appelle fraction *continue*. Pour la fraction continue (1) on utilise également une écriture abrégée

$$a_0+\frac{b_1|}{|a_1}+\frac{b_2|}{|a_2}+\ldots$$

Dans le cas général, les *éléments* d'une fraction continue a_0, a_k, b_k ($k = 1, 2, \ldots$) sont des nombres réels ou complexes ou, encore, les fonctions d'une ou de plusieurs variables. Les fractions $a_0 = \frac{a_0}{1}$, $\frac{b_k}{a_k}$ ($k = 1, 2, \ldots$) s'appellent *termes* d'une fraction continue (1) (respectivement de rang nul, premier, etc.), alors que les nombres ou les fonctions a_k et b_k ($k \geqslant 1$) sont *éléments* du k-ième terme (dénominateurs ou numérateurs partiels). Supposons que $a_k \neq 0$. Constatons que dans l'écriture abrégée (1) les termes $\frac{b_k}{a_k}$ sont irréductibles.

Si une fraction continue (1) compte un nombre fini de termes (n par exemple, sans compter le terme de rang nul), on dit qu'elle est *limitée* ou à n termes, sa notation abrégée est

$$\left[a_0;\frac{b_1}{a_1},\frac{b_2}{a_2},\ldots,\frac{b_n}{a_n}\right]=\left[a_0;\frac{b_k}{a_k}\right]_1^n. \qquad (2)$$

Une fraction continue limitée s'identifie à une fraction ordinaire correspondante obtenue en réalisant les opérations indiquées. Une fraction continue (1) qui possède une infinité de termes est dite *illimitée* et s'écrit

$$\left[a_0;\frac{b_k}{a_k}\right]_1^\infty. \qquad (3)$$

4—01072

La fraction continue

$$a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\ddots}}=\left[a_0;\frac{1}{a_1},\frac{1}{a_2},\ldots\right],\qquad(4)$$

dont tous les numérateurs partiels sont égaux à 1, s'appelle *fraction continue ordinaire* ou *normale*. Les *dénominateurs des termes* s'appellent *quotients incomplets*. Constatons que dans la théorie des nombres les quotients incomplets sont en général des nombres naturels, c'est-à-dire des entiers positifs.

§ 2. Conversion des fractions continues en fractions ordinaires et conversion inverse

Toute fraction continue limitée peut être convertie en fraction ordinaire. A cet effet il suffit de réaliser toutes les opérations indiquées par la notation de la fraction continue.

E x e m p l e 1. Convertir la fraction continue

$$\left[3;\frac{1}{3},\frac{1}{1},\frac{1}{4}\right]=3+\cfrac{1}{3+\cfrac{1}{1+\cfrac{1}{4}}}$$

en fraction ordinaire.

S o l u t i o n. La réalisation successive des opérations imposées conduit à

$$1)\ 1+\frac{1}{4}=\frac{5}{4};\quad 4)\ 1:\frac{19}{5}=\frac{5}{19};$$

$$2)\ 1:\frac{5}{4}=\frac{4}{5};\quad 5)\ 3+\frac{5}{19}=\frac{62}{19}.$$

$$3)\ 3+\frac{4}{5}=\frac{19}{5};$$

Par conséquent,

$$\left[3;\frac{1}{3},\frac{1}{1},\frac{1}{4}\right]=\frac{62}{19}.$$

Inversement, tout nombre rationnel positif peut être converti en fraction continue aux éléments naturels. Soit, par exemple, une fraction $\frac{p}{q}$. En extrayant la partie entière a_0, on a:

$$\frac{p}{q}=a_0+\frac{r_0}{q},$$

où r_0 est le reste $\left(\text{si } \frac{p}{q} \text{ est une fraction propre, } a_0=0 \text{ et } r_0=p\right)$.

Après avoir divisé le numérateur et le dénominateur de la fraction $\frac{r_0}{q}$ par r_0, on obtient:

$$\frac{r_0}{q}=\frac{1}{q:r_0}=\frac{1}{a_1+\frac{r_1}{r_0}},$$

où a_1 est le quotient entier, r_1 le reste de la division de q par r_0.

Après avoir divisé le numérateur et le dénominateur de la fraction $\frac{r_1}{r_0}$ par r_1, on amène

$$\frac{r_1}{r_0}=\frac{1}{r_0:r_1}=\frac{1}{a_2+\frac{r_2}{r_1}},$$

où a_2 est le quotient entier, r_2 le reste de la division de r_0 par r_1. Le processus peut être poursuivi d'une façon analogue.

Puisque $q > r_0 > r_1 > r_2 > r_3 > \ldots$ et r_i $(i = 0, 1, 2, \ldots)$ sont des entiers positifs, on obtient finalement le reste $r_n = 0$, c'est-à-dire

$$\frac{r_{n-1}}{r_{n-2}}=\frac{1}{a_n+0}.$$

En portant l'expression des fractions $\frac{r_i}{r_{i-1}}$, il vient:

$$\frac{p}{q}=a_0+\frac{r_0}{q}=a_0+\frac{1}{a_1+\frac{r_1}{r_0}}=a_0+\frac{1}{a_1+\frac{1}{a_2+\frac{r_2}{r_1}}}=$$

$$=a_0+\frac{1}{a_1+\frac{1}{a_2+\cdot_{\cdot_{\cdot}}+\frac{1}{a_n}}}.$$

Exemple 2. Convertir $\frac{62}{19}$ en fraction continue.

Solution. On a successivement:

$$\frac{62}{19}=3+\frac{5}{19}=3+\frac{1}{\frac{19}{5}}=3+\frac{1}{3+\frac{4}{5}}=3+\frac{1}{3+\frac{1}{\frac{5}{4}}}=3+\frac{1}{3+\frac{1}{1+\frac{1}{4}}}.$$

Ainsi $\frac{62}{19}=\left[3;\frac{1}{3},\frac{1}{1},\frac{1}{4}\right]$.

D'une façon analogue on convertit les fractions continues de forme générale.

Exemple 3. Convertir la fraction continue

$$\left[1;\ \frac{-x^2}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5}\right]=1-\cfrac{x^2}{1-\cfrac{x^2}{3-\cfrac{x^2}{5}}}$$

en fraction ordinaire.

Solution. On a:

$$1)\ 1-\cfrac{x^2}{3-\cfrac{x^2}{5}}=1-\frac{5x^2}{15-x^2}=\frac{15-6x^2}{15-x^2};$$

$$2)\ 1-\cfrac{x^2}{\cfrac{15-6x^2}{15-x^2}}=1-\frac{15x^2-x^4}{15-6x^2}=\frac{15-21x^2+x^4}{15-6x^2}.$$

Ainsi

$$\left[1;\ \frac{-x^2}{1},\ \frac{-x^2}{3},\ \frac{-x^2}{5}\right]=\frac{15-21x^2+x^4}{15-6x^2}.$$

§ 3. Fractions correspondantes

Soit une fonction continue limitée ou illimitée

$$\left[a_0;\ \frac{b_k}{a_k}\right]_1^n. \tag{1}$$

La fraction ordinaire

$$\frac{P_k}{Q_k}\equiv\left[a_0;\ \frac{b_1}{a_1},\ \ldots,\ \frac{b_k}{a_k}\right]$$

($k = 1, 2, \ldots$), où $k \leqslant n$ est dite *k-ième fraction correspondante* de la fraction continue (1). D'après Euler, on adopte généralement

$$\frac{P_0}{Q_0}=\frac{a_0}{1};\qquad \frac{P_{-1}}{Q_{-1}}=\frac{1}{0};$$

et pour écarter l'indétermination, on pose encore

$$P_0 = a_0,\quad Q_0 = 1 \tag{2}$$

et

$$P_{-1} = 1,\quad Q_{-1} = 0. \tag{2'}$$

En travaillant sur un calculateur digital, il est commode de chercher les fractions continues correspondantes

$$\frac{P_n}{Q_n}=a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\ddots+\cfrac{b_n}{a_n}}}$$

à l'aide du *schéma de Hörner* (cf. chapitre III) *pour la division*

$$c_1 = \frac{b_n}{a_n}, \qquad d_1 = a_{n-1} + c_1;$$
$$c_2 = \frac{b_{n-1}}{d_1}, \qquad d_2 = a_{n-2} + c_2;$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$c_k = \frac{b_{n-k+1}}{d_{k-1}}, \qquad d_k = a_{n-k} + c_k;$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$c_n = \frac{b_1}{d_{n-1}}, \qquad d_n = a_0 + c_n = \frac{P_n}{Q_n}.$$

La succession indiquée des opérations se met aisément en programme.

Théorème 1. (Loi de composition des fractions correspondantes.) *Soient les nombres* P_k, Q_k $(k = 1, 2, \ldots)$ *définis par les relations*

$$P_k = a_k P_{k-1} + b_k P_{k-2}, \tag{3}$$
$$Q_k = a_k Q_{k-1} + b_k Q_{k-2} \tag{3'}$$

avec

$$P_{-1} = 1, \quad Q_{-1} = 0; \quad P_0 = a_0, \quad Q_0 = 1. \tag{4}$$

Alors les fractions $\frac{P_k}{Q_k}$ *aux termes ainsi définis sont des fractions correspondantes de la fraction continue* (1) *.

Démonstration. Soit R_k $(k = 1, 2, \ldots)$ les fractions correspondantes successives de la fraction continue (1). Montrer que

$$R_k = \frac{P_k}{Q_k} \quad (k = 1, 2, \ldots).$$

La démonstration se fait par récurrence.

Avec $k = 1$ on a pour la fraction correspondante R_1

$$R_1 = a_0 + \frac{b_1}{a_1} = \frac{a_0 a_1 + b_1}{a_1}.$$

Par ailleurs, tenant compte de (4), les relations (3) et (3') entraînent

$$P_1 = a_1 a_0 + b_1,$$
$$Q_1 = a_1 \cdot 1 + b_1 \cdot 0 = a_1.$$

Par conséquent, $R_1 = \frac{P_1}{Q_1}$ et pour $k = 1$ le théorème est vérifié.

---

* Les fractions correspondantes aux termes ainsi définis sont dites *canoniques*.

Supposons maintenant que le théorème soit vrai pour tout nombre naturel inférieur ou égal à k. Montrons que le théorème est valable également pour le nombre naturel successif $k+1$. Des relations (3) et (3′) il suit que

$$P_{k+1} = a_{k+1}P_k + b_{k+1}P_{k-1},$$
$$Q_{k+1} = a_{k+1}Q_k + b_{k+1}Q_{k-1}.$$

Par hypothèse

$$R_k = \frac{P_k}{Q_k} = \frac{a_kP_{k-1}+b_kP_{k-2}}{a_kQ_{k-1}+b_kQ_{k-2}}$$

D'après la méthode de composition d'une fraction continue (1), la fraction correspondante R_{k+1} s'obtient à partir de R_k en remplaçant l'élément a_k par la somme $a_k + \frac{b_{k+1}}{a_{k+1}}$. C'est pourquoi

$$R_{k+1} = \frac{\left(a_k+\frac{b_{k+1}}{a_{k+1}}\right)P_{k-1}+b_kP_{k-2}}{\left(a_k+\frac{b_{k+1}}{a_{k+1}}\right)Q_{k-1}+b_kQ_{k-2}} = \frac{a_{k+1}(a_kP_{k-1}+b_kP_{k-2})+b_{k+1}P_{k-1}}{a_{k+1}(a_kQ_{k-1}+b_kQ_{k-2})+b_{k+1}Q_{k-1}} =$$
$$= \frac{a_{k+1}P_k+b_{k+1}P_{k-1}}{a_{k+1}Q_k+b_{k+1}Q_{k-1}} = \frac{P_{k+1}}{Q_{k+1}},$$

ce qu'il fallait démontrer.

R e m a r q u e. La détermination des termes des fractions correspondantes étant non univoque, on ne peut affirmer dans le cas général que le numérateur et le dénominateur des fractions correspondantes non canoniques vérifient les équations (3) et (3′). Par la suite nous supposerons que les fractions correspondantes considérées sont canoniques.

C o r o l l a i r e. Pour une fraction continue ordinaire

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}$$

les numérateurs et les dénominateurs de ses fractions correspondantes $\frac{p_k}{q_k}$ ($k = 1, 2, \ldots$) peuvent être déterminés à partir des relations

$$\left.\begin{aligned} p_k &= a_kp_{k-1} + p_{k-2},\\ q_k &= a_kq_{k-1} + q_{k-2},\end{aligned}\right\} \qquad (3'')$$

où l'on a posé $p_0 = a_0$, $p_{-1} = 1$ et $q_0 = 1$, $q_{-1} = 0$.

R e m a r q u e. Pour chercher d'après les formules (3) et (3′) les fractions correspondantes successives il est commode d'utiliser le schéma suivant:

k	-1	0	1	2	3	...
b_k		1	b_1	b_2	b_3	...
a_k		a_0	a_1	a_2	a_3	...
P_k	1	a_0	P_1	P_2	P_3	...
Q_k	0	1	Q_1	Q_2	Q_3	...

Pour une fraction continue ordinaire où $b_k = 1$ ($k = 1, 2, \ldots$) et qui donne lieu aux formules (3″) on élimine du schéma la ligne b_k.

Exemple 1. Calculer toutes les fractions correspondantes de la fraction continue

$$\frac{163}{59} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{4 + \cfrac{1}{1 + \cfrac{1}{2}}}}}.$$

Solution. Appliquons le schéma ci-dessus pour obtenir

a_k		2	1	3	4	1	2
p_k	$p_{-1} = 1$	2	3	11	47	58	163
q_k	$q_{-1} = 0$	1	1	4	17	21	59

Par conséquent,

$$\frac{p_0}{q_0} = \frac{2}{1}; \quad \frac{p_1}{q_1} = \frac{3}{1}; \quad \frac{p_2}{q_2} = \frac{11}{4};$$

$$\frac{p_3}{q_3} = \frac{47}{17}; \quad \frac{p_4}{q_4} = \frac{58}{21}; \quad \frac{p_5}{q_5} = \frac{163}{59}.$$

Exemple 2. Trouver toutes les fractions correspondantes de la fraction continue

$$\left[0; \frac{1}{2}, \frac{3}{4}, \frac{5}{8}, \frac{7}{16}\right].$$

Solution. En appliquant le schéma donné plus haut on a

k	-1	0	1	2	3	4
b_k		1	1	3	5	7
a_k		0	2	4	8	16
P_k	1	0	1	4	37	620
Q_k	0	1	2	11	98	1645

Il s'ensuit que

$$\frac{P_0}{Q_0}=\frac{0}{1};\quad \frac{P_1}{Q_1}=\frac{1}{2};\quad \frac{P_2}{Q_2}=\frac{4}{11};\quad \frac{P_3}{Q_3}=\frac{37}{98};\quad \frac{P_4}{Q_4}=\frac{620}{1645}.$$

Théorème 2. *Deux fractions correspondantes voisines* $\frac{P_{k-1}}{Q_{k-1}}$ *et* $\frac{P_k}{Q_k}$ *de la fraction continue* (1) *vérifient la formule*

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}\qquad (k\geqslant 1). \tag{4'}$$

Démonstration. On a:

$$\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}=\frac{\Delta_k}{Q_{k-1}Q_k}, \tag{5}$$

où

$$\Delta_k=\begin{vmatrix} P_k & P_{k-1}\\ Q_k & Q_{k-1}\end{vmatrix}.$$

En utilisant les relations (3) et (3′) on obtient, en vertu des propriétés connues du déterminant,

$$\Delta_k=\begin{vmatrix} a_kP_{k-1}+b_kP_{k-2} & P_{k-1}\\ a_kQ_{k-1}+b_kQ_{k-2} & Q_{k-1}\end{vmatrix}=b_k\begin{vmatrix} P_{k-2} & P_{k-1}\\ Q_{k-2} & Q_{k-1}\end{vmatrix}=-b_k\Delta_{k-1}.$$

On en tire successivement:

$$\Delta_k=(-b_k)(-b_{k-1})\ldots(-b_1)\Delta_0=(-1)^k b_1b_2\ldots b_k\Delta_0,$$

où

$$\Delta_0=\begin{vmatrix} P_0 & P_{-1}\\ Q_0 & Q_{-1}\end{vmatrix}=\begin{vmatrix} a_0 & 1\\ 1 & 0\end{vmatrix}=-1.$$

Donc

$$\Delta_k = (-1)^{k-1} b_1 b_2 \ldots b_k,$$

et, par conséquent, la formule (5) permet de déduire

$$\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}} = (-1)^{k-1} \frac{b_1 b_2 \ldots b_k}{Q_{k-1} Q_k}.$$

Corollaire 1. Si $\frac{P_{k-1}}{Q_{k-1}}$ et $\frac{P_k}{Q_k}$ $(k \geqslant 1)$ sont deux fractions correspondantes voisines de la fraction continue (1),

$$\Delta_k = P_k Q_{k-1} - P_{k-1} Q_k = (-1)^{k-1} b_1 b_2 \ldots b_k.$$

Corollaire 2. Deux fractions correspondantes voisines $\frac{p_{k-1}}{q_{k-1}}$, $\frac{p_k}{q_k}$ $(k \geqslant 1)$ d'une fraction continue ordinaire vérifient l'égalité

$$\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{(-1)^{k-1}}{q_{k-1} q_k}. \qquad (4'')$$

Théorème 3. *Deux fractions correspondantes voisines de même parité de rang* $\frac{P_{k-2}}{Q_{k-2}}$ *et* $\frac{P_k}{Q_k}$ $(k \geqslant 2)$ *de la fraction continue* (1) *vérifient la relation*

$$\frac{P_k}{Q_k} - \frac{P_{k-2}}{Q_{k-2}} = (-1)^k \frac{b_1 b_2 \ldots b_{k-1} a_k}{Q_{k-2} Q_k}. \qquad (6)$$

Démonstration. On a

$$\frac{P_k}{Q_k} - \frac{P_{k-2}}{Q_{k-2}} = \frac{D_k}{Q_{k-2} Q_k}, \qquad (7)$$

où

$$D_k = \begin{vmatrix} P_k & P_{k-2} \\ Q_k & Q_{k-2} \end{vmatrix}.$$

On en déduit en vertu de la loi de composition des fractions correspondantes et des propriétés élémentaires du déterminant

$$D_k = \begin{vmatrix} a_k P_{k-1} + b_k P_{k-2} & P_{k-2} \\ a_k Q_{k-1} + b_k Q_{k-2} & Q_{k-2} \end{vmatrix} = a_k \begin{vmatrix} P_{k-1} & P_{k-2} \\ Q_{k-1} & Q_{k-2} \end{vmatrix} = a_k \Delta_{k-1},$$

où Δ_k est le déterminant examiné dans le théorème 2. D'après le corollaire 1 du théorème 1

$$\Delta_{k-1} = (-1)^k b_1 b_2 \ldots b_{k-1},$$

d'où

$$D_k = (-1)^k b_1 b_2 \ldots b_{k-1} a_k.$$

Par conséquent, en appliquant la relation (7) on obtient la formule (6).

Corollaire. Si $\frac{p_{k-2}}{q_{k-2}}$ et $\frac{p_k}{q_k}$ sont deux fractions correspondantes voisines de même parité d'une fraction continue ordinaire

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}},$$

elles donnent lieu à la relation

$$\frac{p_k}{q_k} - \frac{p_{k-2}}{q_{k-2}} = (-1)^k \frac{a_k}{q_{k-2}q_k}. \tag{6'}$$

Théorème 4. *Si tout élément d'une fraction continue limitée est positif, ses fractions correspondantes de rang pair forment une suite croissante, alors que ses fractions correspondantes de rang impair forment une suite décroissante. Par ailleurs, toute fraction correspondante de rang pair est inférieure à toute fraction correspondante de rang impair. Quant au nombre α exprimé par la fraction continue, il est compris entre deux fractions correspondantes voisines.*

Démonstration. Soit la fraction continue

$$\alpha = \left[a_0; \frac{b_1}{a_1}, \frac{b_2}{a_2}, \ldots, \frac{b_n}{a_n}\right] \tag{8}$$

à éléments positifs a_k et b_k et soit $\frac{P_k}{Q_k}$ $(k = 0, 1, \ldots, n)$ ses fractions correspondantes canoniques successives. Il est clair que $P_k > 0$ et $Q_k > 0$.

Considérons deux cas.

1. Soit $k = 2m$ un nombre pair. La relation (6) entraîne alors compte tenu du fait que $a_k > 0$ et $b_i > 0$ $(i = 1, \ldots, k)$:

$$\frac{P_{2m}}{Q_{2m}} - \frac{P_{2m-2}}{Q_{2m-2}} > 0.$$

Par conséquent,

$$\frac{P_{2m-2}}{Q_{2m-2}} < \frac{P_{2m}}{Q_{2m}} \quad (m = 1, 2, \ldots)$$

ou

$$\frac{P_0}{Q_0} < \frac{P_2}{Q_2} < \frac{P_4}{Q_4} < \ldots \tag{9}$$

2. Soit $k = 2m + 1$ un nombre impair. Donc, $k - 1$ sera un nombre pair. Alors la même relation (6) amène

$$\frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m+1}}{Q_{2m+1}}$$

ou

$$\frac{P_1}{Q_1} > \frac{P_3}{Q_3} > \frac{P_5}{Q_5} > \dots \tag{10}$$

On a montré ainsi que les fractions correspondantes de rang pair forment une suite croissante et celles de rang impair une suite décroissante (fig. 1).

Ensuite, si dans la relation (4′) on pose $k = 2m$, il vient :

$$\frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m}}{Q_{2m}}, \tag{11}$$

c'est-à-dire toute fraction correspondante de rang impair est plus grande qu'une fraction croissante suivante de rang pair. On en

Fig. 1.

déduit qu'une fraction correspondante quelconque de rang impair est plus grande qu'une fraction croissante quelconque de rang pair. En effet, soit $\frac{P_{2s-1}}{Q_{2s-1}}$ une quelconque fraction correspondante de rang impair. Si $s \leqslant m$,

$$\frac{P_{2s-1}}{Q_{2s-1}} \geqslant \frac{P_{2m-1}}{Q_{2m-1}} > \frac{P_{2m}}{Q_{2m}};$$

si $s > m$,

$$\frac{P_{2s-1}}{Q_{2s-1}} > \frac{P_{2s}}{Q_{2s}} > \frac{P_{2m}}{Q_{2m}}.$$

Par conséquent, quels que soient s et m, on a :

$$\frac{P_{2s-1}}{Q_{2s-1}} > \frac{P_{2m}}{Q_{2m}}. \tag{12}$$

Enfin, le mode de composition de la fraction continue

$$\alpha = a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \ddots}}$$

conduit aux relations évidentes

$$\alpha > \frac{P_0}{Q_0}, \quad \alpha < \frac{P_1}{Q_1}, \quad \alpha > \frac{P_2}{Q_2}, \dots$$

Donc,

$$\frac{P_k}{Q_k} < \alpha < \frac{P_{k+1}}{Q_{k+1}} \tag{13}$$

si k est un nombre pair, et

$$\frac{P_k}{Q_k} > \alpha > \frac{P_{k+1}}{Q_{k+1}} \tag{13'}$$

si k est un nombre impair. Il est clair que la dernière fraction correspondante donne lieu non pas aux inégalités strictes (13) et (13′) mais à une égalité à droite.

Corollaire 1. Si les éléments d'une fraction continue (8) sont positifs et $\frac{P_k}{Q_k}$ sont ses fractions correspondantes, l'estimation

$$\left|\alpha - \frac{P_k}{Q_k}\right| \leqslant \frac{b_1 b_2 \ldots b_{k+1}}{Q_k Q_{k+1}} \tag{14}$$

se trouve vérifiée.

En effet, puisque, d'après ce qui a été démontré,

$$\left|\alpha - \frac{P_k}{Q_k}\right| \leqslant \left|\frac{P_{k+1}}{Q_{k+1}} - \frac{P_k}{Q_k}\right|,$$

la formule (4′) entraîne l'estimation (14).

Corollaire 2. Si la fraction continue α aux éléments positifs est ordinaire et $\frac{p_k}{q_k}$ sont ses fractions correspondantes successives,

$$\left|\alpha - \frac{p_k}{q_k}\right| \leqslant \frac{1}{q_k q_{k+1}}.$$

Remarque. Si les éléments d'une fraction continue ordinaire sont des nombres naturels, on peut montrer [1] que la fraction correspondante $\frac{p_k}{q_k}$ est la *meilleure approximation* du nombre α, c'est-à-dire que toutes les autres fractions $\frac{p}{q}$ au dénominateur $q \leqslant q_k$ s'écartent du nombre α plus que la fraction $\frac{p_k}{q_k}$.

Exemple 3. La fraction $\frac{58}{21}$ était l'avant-dernière fraction correspondante de $\frac{163}{59}$ (cf. exemple 1). Donc

$$\left|\frac{163}{59} - \frac{58}{21}\right| \leqslant \frac{1}{59 \cdot 21} < 0{,}001.$$

§ 4. Fractions continues illimitées

Soit

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots\right]=a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\ddots}} \tag{1}$$

une fraction continue illimitée. Considérons un de ses segments, c'est-à-dire une fraction continue limitée

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n}\right]=\frac{P_n}{Q_n} \quad (n=1,\ 2,\ 3,\ \ldots). \tag{2}$$

D é f i n i t i o n. Une fraction continue illimitée (1) s'appelle convergente s'il existe une limite finie

$$\alpha=\lim_{n\to\infty}\frac{P_n}{Q_n}, \tag{3}$$

le nombre α étant considéré comme une valeur de cette fraction. Si la limite (3) n'existe pas, la fraction continue (1) est dite *divergente* et on ne lui affecte aucune valeur numérique.

D'après le *critère de Cauchy* [3] pour rendre la suite $\frac{P_n}{Q_n}$ $(n = 1, 2, 3, \ldots)$ convergente, il faut et il suffit que, pour tout $\varepsilon > 0$, il existe un nombre $N = N(\varepsilon)$ tel que

$$\left|\frac{P_{n+m}}{Q_{n+m}}-\frac{P_n}{Q_n}\right|<\varepsilon$$

avec $n > N$ et n'importe quel $m > 0$.

Si $Q_k \neq 0$, on a évidemment:

$$\frac{P_n}{Q_n}=\frac{P_0}{Q_0}+\sum_{k=1}^{n}\left(\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right). \tag{4}$$

D'où

$$\lim_{n\to\infty}\frac{P_n}{Q_n}=\frac{P_0}{Q_0}+\sum_{k=1}^{\infty}\left(\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right)=\frac{P_0}{Q_0}+\sum_{k=2}^{\infty}(-1)^{k-1}\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}, \tag{4'}$$

c'est-à-dire la convergence de la fraction continue (1) est équivalente à la convergence de la série (4′). Si la fraction continue (1) converge:

$$\alpha=\lim_{n\to\infty}\frac{P_n}{Q_n},$$

les formules (4) et (4′) donnent l'estimation

$$\left|\alpha-\frac{P_n}{Q_n}\right|\leqslant\sum_{k=n+1}^{\infty}\left|\frac{P_k}{Q_k}-\frac{P_{k-1}}{Q_{k-1}}\right|\leqslant\sum_{k=n+1}^{\infty}\left|\frac{b_1b_2\ldots b_k}{Q_{k-1}Q_k}\right|.$$

Théorème 1. *Si tous les éléments* a_k, b_k $(k = 0, 1, \ldots)$ *d'une fraction continue* (1) *sont positifs, et en outre*

$$b_k \leqslant a_k \quad \text{et} \quad a_k \geqslant d > 0 \quad (k = 1, 2, \ldots), \tag{5}$$

la fraction continue (1) *est convergente.*

Démonstration. Dans la démonstration de la première partie du théorème 4 du paragraphe précédent nous n'avons pas utilisé le fait que la fraction continue était limitée. C'est pourquoi, en reprenant cette démonstration, on établit que si tous les éléments d une fraction continue (1) sont positifs, ses fractions correspondantes de rang pair $\frac{P_{2k}}{Q_{2k}}$ $(k = 0, 1, 2, \ldots)$ forment une suite croissante bornée supérieurement (par exemple, par le nombre $\frac{P_1}{Q_1}$). On en tire, en raison du théorème connu, qu'il existe une limite

$$\lim_{k\to\infty} \frac{P_{2k}}{Q_{2k}} = \alpha.$$

Sous les conditions imposées par le théorème ci-dessus, les fractions correspondantes de rang impair $\frac{P_{2k+1}}{Q_{2k+1}}$ $(k = 0, 1, 2, \ldots)$ de la fraction continue (1) forment d'une façon analogue une suite décroissante bornée inférieurement $\left(\text{par le nombre } \frac{P_0}{Q_0}, \text{ par exemple}\right)$. Il existe donc encore une limite

$$\lim_{k\to\infty} \frac{P_{2k+1}}{Q_{2k+1}} = \beta$$

et de plus, $\beta \geqslant \alpha$. Par ailleurs, pour tout $k \geqslant 0$ on a:

$$\frac{P_{2k}}{Q_{2k}} < \alpha \leqslant \beta < \frac{P_{2k+1}}{Q_{2k+1}};$$

c'est pourquoi en appliquant le théorème 2 du § 3 on obtient

$$0 \leqslant \beta - \alpha < \frac{P_{2k+1}}{Q_{2k+1}} - \frac{P_{2k}}{Q_{2k}} = \frac{b_1 b_2 \ldots b_{2k+1}}{Q_{2k} Q_{2k+1}} = \eta_k. \tag{6}$$

Montrons que $\eta_k \to 0$ lorsque $k \to \infty$. En effet, suivant la loi de composition des fractions correspondantes on a avec $k \geqslant 2$:

$$Q_k = a_k Q_{k-1} + b_k Q_{k-2}$$

et

$$Q_{k-1} = a_{k-1} Q_{k-2} + b_{k-1} Q_{k-3}.$$

Les conditions (5) du théorème permettent de déduire

$$Q_k \geqslant b_k (Q_{k-1} + Q_{k-2})$$

et

$$Q_{k-1} \geqslant d Q_{k-2}.$$

Donc,

$$Q_k \geqslant b_k (1 + d) Q_{k-2}. \qquad (7)$$

L'inégalité (7) donne successivement

$$Q_{2k} \geqslant b_{2k} (1 + d) Q_{2k-2} \geqslant \ldots \geqslant b_{2k} b_{2k-2} \ldots b_2 (1 + d)^k Q_0 = \\ = b_2 b_4 \ldots b_{2k} (1 + d)^k \qquad (8)$$

et

$$Q_{2k+1} \geqslant b_{2k+1} (1 + d) Q_{2k-1} \geqslant \ldots \\ \ldots \geqslant b_{2k+1} \ldots b_3 (1 + d)^k Q_1 \geqslant b_1 b_3 \ldots b_{2k+1} (1 + d)^k, \qquad (9)$$

puisque $Q_1 = a_1 \geqslant b_1$. En multipliant les inégalités (8) et (9) on tombe sur

$$Q_{2k} Q_{2k+1} \geqslant b_1 b_2 \ldots b_{2k+1} (1 + d)^{2k} \qquad (10)$$

et, par conséquent,

$$\eta_k = \frac{b_1 b_2 \ldots b_{2k+1}}{Q_{2k} Q_{2k+1}} \leqslant \frac{1}{(1+d)^{2k}}.$$

Ainsi, $\eta_k \to 0$ lorsque $k \to \infty$.

C'est pourquoi en passant à la limite lorsque $k \to \infty$ on a dans l'inégalité (6) $0 \leqslant \beta - \alpha \leqslant 0$, c'est-à-dire

$$\alpha = \beta = \lim_{n \to \infty} \frac{P_n}{Q_n},$$

et donc la fraction continue (1) est convergente.

R e m a r q u e. Pour la fraction convergente (1) à éléments positifs sa valeur α est comprise entre deux fractions correspondantes consécutives $\frac{P_{n-1}}{Q_{n-1}}$ et $\frac{P_n}{Q_n}$. Il s'ensuit que

$$\left| \alpha - \frac{P_n}{Q_n} \right| \leqslant \left| \frac{P_n}{Q_n} - \frac{P_{n-1}}{Q_{n-1}} \right| = \frac{b_1 b_2 \ldots b_n}{Q_{n-1} Q_n}.$$

C o r o l l a i r e. Une fraction continue ordinaire à éléments naturels est toujours convergente.

On peut démontrer également le théorème suivant [1].

T h é o r è m e 2. *Tout nombre positif α peut être développé en fraction continue convergente ordinaire à éléments naturels, ce développement étant unique.* La fraction continue ainsi obtenue est limitée si α est un nombre rationnel, elle est illimitée si α est un nombre irrationnel.

E x e m p l e. Développer en fraction continue le nombre $\sqrt{41}$ et trouver sa valeur approchée.

S o l u t i o n. L'entier maximal contenu dans $\sqrt{41}$ étant 6, on a:

$$\sqrt{41} = 6 + \frac{1}{a_1}. \qquad (11)$$

Il en résulte

$$a_1=\frac{1}{\sqrt{41}-6}=\frac{6+\sqrt{41}}{5}.$$

L'entier maximal contenu dans a_1 est 2, c'est pourquoi

$$a_1=2+\frac{1}{a_2}. \tag{12}$$

Ce qui amène

$$a_2=\frac{1}{a_1-2}=\frac{5}{\sqrt{41}-4}=\frac{4+\sqrt{41}}{5}=2+\frac{1}{a_3}. \tag{13}$$

D'une façon analogue

$$a_3=\frac{1}{a_2-2}=\frac{5}{\sqrt{41}-6}=6+\sqrt{41}=12+\frac{1}{a_4}; \tag{14}$$

$$a_4=\frac{1}{a_3-12}=\frac{1}{\sqrt{41}-6}=\frac{6+\sqrt{41}}{5}=2+\frac{1}{a_5}. \tag{15}$$

Nous constatons que $a_4 = a_1$, les éléments de la fraction continue reviennent donc, et c'est pourquoi la fraction obtenue est périodique. En portant successivement dans l'égalité (11) les expressions (12), (13), (14), (15), etc., on obtient:

$$\sqrt{41}=6+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{12+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{12+\ddots}}}}}}$$

Le nombre irrationnel $\sqrt{41}$ s'exprime ainsi par une fraction continue périodique illimitée

$$\sqrt{41}=\left(6;\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{12},\ \ldots\right).$$

Les fractions correspondantes $\frac{p_k}{q_k}$ ($k = 0, 1, 2, \ldots$) s'obtiennent en utilisant le schéma suivant:

a_k	—	6	2	2	12	2	2	12
p_k	$p_{-1}=1$	$p_0=6$	13	32	397	826	2049	...
q_k	$q_{-1}=0$	$q_0=1$	2	5	62	129	320	...

En se bornant, par exemple, à la cinquième fraction correspondante, on obtient la valeur approchée par excès: $\sqrt{41} = \frac{2049}{320} = 6{,}403125$ avec une erreur absolue plus petite que

$$\Delta = \frac{1}{320\,(2\cdot 320 + 129)} = \frac{1}{320\cdot 769} < 5\cdot 10^{-6}.$$

Théorème 3 (de Pringsheim). *Si une fraction continue illimitée*

$$\left[0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \ldots,\ \frac{b_n}{a_n}\right] \tag{16}$$

vérifie les inégalités

$$|b_n| + 1 \leqslant |a_n| \quad (n = 1,\ 2,\ \ldots), \tag{17}$$

cette fraction est convergente et en outre sa valeur absolue ne dépasse pas l'unité [4].

Démonstration. Soient $\frac{P_k}{Q_k}$ $(k = 1,\ 2,\ \ldots)$ les fractions correspondantes de la fraction continue (16).

Etant donné que

$$Q_k = a_k Q_{k-1} + b_k Q_{k-2} \quad (k = 1,\ 2,\ \ldots),$$

il vient

$$|Q_k| \geqslant |a_k|\,|Q_{k-1}| - |b_k|\,|Q_{k-2}|.$$

On en tire en appliquant l'inégalité (17):

$$|Q_k| \geqslant (|b_k| + 1)\,|Q_{k-1}| - |b_k|\,|Q_{k-2}|,$$

ou

$$|Q_k| - |Q_{k-1}| \geqslant |b_k|\,(|Q_{k-1}| - |Q_{k-2}|). \tag{18}$$

En utilisant successivement l'inégalité (18) et en retenant que $Q_0 = 1$ et $Q_{-1} = 0$, on a:

$$|Q_k| - |Q_{k-1}| \geqslant |b_k|\,|b_{k-1}| \ldots |b_1|. \tag{19}$$

L'inégalité (19) entraîne que $|Q_k|$ croît monotonement avec l'augmentation de k, de plus $|Q_k| \geqslant |Q_0| = 1$.

La convergence de la fraction continue (16) est équivalente à la convergence de la série

$$\frac{P_0}{Q_0} + \sum_{k=1}^{\infty} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}\, b_1 b_2 \ldots b_k}{Q_{k-1} Q_k}. \tag{20}$$

Considérons la série des modules

$$\sum_{k=1}^{\infty} \frac{|b_1|\,|b_2| \ldots |b_k|}{|Q_{k-1}|\,|Q_k|}. \tag{21}$$

L'inégalité (19) donne:

$$\sum_{k=1}^{n} \frac{|b_1||b_2|\ldots|b_k|}{|Q_{k-1}||Q_k|} \leqslant \sum_{k=1}^{n} \frac{|Q_k|-|Q_{k-1}|}{|Q_{k-1}||Q_k|} =$$

$$= \sum_{k=1}^{n} \left(\frac{1}{|Q_{k-1}|} - \frac{1}{|Q_k|}\right) = \frac{1}{|Q_0|} - \frac{1}{|Q_n|} < \frac{1}{|Q_0|} = 1 \qquad (n = 1, 2, \ldots).$$

Ainsi les sommes partielles de la série (21) sont bornées et, par conséquent, cette série est convergente, en outre

$$\sum_{k=1}^{\infty} \frac{|b_1||b_2|\ldots|b_k|}{|Q_{k-1}||Q_k|} \leqslant 1. \tag{22}$$

Mais alors, en vertu du critère de comparaison, la série (20) converge elle aussi et cette convergence est absolue, ce qui veut dire qu'il existe une limite

$$\lim_{n\to\infty} \frac{P_n}{Q_n} = \sum_{k=1}^{\infty} \left(\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right) = \alpha.$$

Par ailleurs, compte tenu de l'inégalité (22), on a:

$$|\alpha| \leqslant 1.$$

Remarque 1. Pour que la fraction continue (16) converge, il suffit que l'inégalité (17) soit satisfaite pour $n \geqslant m$ et que $Q_k \neq 0$ pour $k \leqslant m$.

Remarque 2. Sous les conditions du théorème 3, nous avons l'estimation suivante de la valeur de la fraction continue α:

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \sum_{k=n+1}^{\infty} \left|\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}}\right| =$$

$$= \sum_{k=n+1}^{\infty} \frac{|b_1||b_2|\ldots|b_k|}{|Q_{k-1}||Q_k|} \leqslant \sum_{k=n+1}^{\infty} \frac{|Q_k|-|Q_{k-1}|}{|Q_{k-1}||Q_k|} =$$

$$= \sum_{k=n+1}^{\infty} \left(\frac{1}{|Q_{k-1}|} - \frac{1}{|Q_k|}\right) = \frac{1}{|Q_n|} - \lim_{k\to\infty} \frac{1}{|Q_k|}.$$

En particulier, si $|Q_k| \to +\infty$ lorsque $k \to \infty$, on a

$$\left|\alpha - \frac{P_n}{Q_n}\right| \leqslant \frac{1}{|Q_n|}.$$

§ 5. Développement des fonctions en fractions continues

Les fractions continues sont un outil commode pour la représentation et le calcul des fonctions. Cet aspect de la question est examiné en détail dans des ouvrages appropriés (cf., par exemple, [2]), alors que nous ne nous bornerons qu'à l'étude de cas particuliers.

Constatons que si nous recourons à un artifice quelconque pour développer la fonction $f(x)$ en une fraction continue illimitée, dans le cas général il faut montrer la convergence de cette fraction et voir si la valeur limite de la fraction continue est égale à la fonction $f(x)$.

A. Développement d'une fonction rationnelle en fraction continue

Si

$$f(x)=\frac{c_{10}+c_{11}x+c_{12}x^2+\ldots}{c_{00}+c_{01}x+c_{02}x^2+\ldots},$$

en effectuant des transformations élémentaires, on aboutit dans le cas général à

$$f(x)=\frac{1}{\frac{c_{00}}{c_{10}}+\frac{c_{00}+c_{01}x+c_{02}x^2+\ldots}{c_{10}+c_{11}x+c_{12}x^2+\ldots}-\frac{c_{00}}{c_{10}}}=\frac{c_{10}}{c_{00}+xf_1(x)},$$

où

$$f_1(x)=\frac{c_{20}+c_{21}x+c_{22}x^2+\ldots}{c_{10}+c_{11}x+c_{12}x^2+\ldots}$$

et

$$c_{2k}=c_{10}c_{0,\,k+1}-c_{00}c_{1,\,k+1}\quad (k=0,\ 1,\ \ldots).$$

D'une façon analogue

$$f_1(x)=\frac{c_{20}}{c_{10}+xf_2(x)},$$

où

$$f_2(x)=\frac{c_{30}+c_{31}x+c_{32}x^2+\ldots}{c_{20}+c_{21}x+c_{22}x^2+\ldots}$$

et

$$c_{3k}=c_{20}c_{1,\,k+1}-c_{10}c_{2,\,k+1}\quad (k=0,\ 1,\ \ldots),$$

etc.

Ainsi

$$f(x)=\cfrac{c_{10}}{c_{00}+\cfrac{c_{20}x}{c_{10}+\cfrac{c_{30}x}{c_{20}+\ddots}}}=\left[0;\ \frac{c_{10}}{c_{00}},\ \frac{c_{20}x}{c_{10}},\ \frac{c_{30}x}{c_{20}},\ \ldots,\ \frac{c_{n0}x}{c_{n-1,\,0}}\right],\qquad (1)$$

et on voit aisément que la fraction continue (1) obtenue est limitée.

Pour calculer successivement les coefficients des développements c_{jk}, il est commode de recourir à la formule

$$c_{jk} = -\begin{vmatrix} c_{j-2,0} & c_{j-2,k+1} \\ c_{j-1,0} & c_{j-1,k+1} \end{vmatrix},$$

où $j \geqslant 2$.

Remarquons que dans certains cas, les coefficients c_{jk} peuvent être nuls. Il faut alors apporter au développement (1) des modifications appropriées [2].

Exemple 1. Développer en fraction continue la fonction

$$f(x) = \frac{1-x}{1-5x+6x^2}.$$

Solution. Portons les coefficients c_{jk} sur le schéma suivant:

j \ k	0	1	2
0	1	−5	6
1	1	−1	0
2	−4	6	0
3	−2	0	0
4	−12	0	0

Par conséquent,

$$\frac{1-x}{1-5x+6x^2} = \left[0;\ \frac{1}{1},\ \frac{-4x}{1},\ \frac{-2x}{-4},\ \frac{-12x}{-2}\right] = \cfrac{1}{1-\cfrac{4x}{1-\cfrac{2x}{-4+6x}}}.$$

B. Développement de e^x en fraction continue

Euler a obtenu pour e^x le développement [2]

$$e^x = \left[0;\ \frac{1}{1},\ \frac{-2x}{2+x},\ \frac{x^2}{6},\ \frac{x^2}{10},\ \ldots,\ \frac{x^2}{4n+2},\ \ldots\right] \tag{2}$$

convergent pour toute valeur de x, réelle ou complexe [2].

On en tire des fractions correspondantes:

$$\frac{P_1}{Q_1} = \frac{1}{1};$$

$$\frac{P_2}{Q_2} = \frac{2+x}{2-x};$$

$$\frac{P_3}{Q_3} = \frac{12+6x+x^2}{12-6x+x^2};$$

$$\frac{P_4}{Q_4} = \frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3},$$

etc.

En posant notamment $x = 1$ et en se bornant à la quatrième fraction correspondante, on a

$$e \approx \frac{193}{71} = 2,7183 \ldots$$

Pour obtenir la même précision, il faut prendre dans le développement de Maclaurin

$$e = 2 + \frac{1}{2!} + \frac{1}{3!} + \ldots$$

au moins huit termes.

C. Développement de tg x en fraction continue

Pour tg x Lambert a obtenu le développement

$$\operatorname{tg} x = \left[0; \frac{x}{1}, \frac{-x^2}{3}, \frac{-x^2}{5}, \ldots, \frac{-x^2}{2n+1}, \ldots\right] \qquad (3)$$

convergent en tout point où la fonction est continue.

Exemple 2. Chercher la valeur approchée de tg 1.

Solution. En posant dans la formule (3) $x = 1$, on a:

$$\operatorname{tg} 1 = \left[0; \frac{1}{1}, \frac{-1}{3}, \frac{-1}{5}, \ldots\right].$$

La formule (3) du § 3 permet de composer le schéma suivant pour les termes des fractions correspondantes:

k	-1	0	1	2	3	4
b_k		1	1	-1	-1	1
a_k		0	1	3	5	7
P_k	1	0	1	3	14	95
Q_k	0	1	1	2	9	61

En se bornant à la quatrième fraction correspondante, on a

$$\operatorname{tg} 1 \approx \frac{95}{61} = 1,557377$$

(la valeur tabulaire est tg 1 = 1,557396).

BIBLIOGRAPHIE

1. *A. Khintchine.* Fractions continues. Gostekhizdat, 1949, chapitre I.
2. *A. Khovanski.* Applications des fractions continues et de leurs généralisations aux problèmes de l'analyse approchée. Gostekhizdat, 1956, chapitres I et II.
3. *G. Fichtengoltz.* Eléments d'analyse mathématique, t. 1. Gostekhizdat, 1955, chapitre III.
4. *O. Perron.* Die Lehre von den Kettenbrüchen. Teubner, 1913, chapitre VII.

CHAPITRE III

CALCUL DES VALEURS DES FONCTIONS

Pour calculer sur des ordinateurs les valeurs des fonctions données par des formules, la forme de ces dernières est loin d'être indifférente. Considérées sous l'optique des calculs approchés, les expressions mathématiques équivalentes n'ont pas toujours la même valeur. Il se pose donc un problème de grand intérêt pratique qui consiste à rechercher pour les fonctions élémentaires les expressions analytiques les plus commodes. Le calcul des valeurs des fonctions se ramène en général à réaliser une suite d'opérations arithmétiques élémentaires. Le volume de la mémoire d'une machine étant limité, il convient de diviser ces opérations en cycles répétitifs. Dans ce qui suit nous allons étudier certains procédés types.

§ 1. Valeurs d'un polynôme. Schéma de Hörner

Soit un polynôme de degré n

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \tag{1}$$

à coefficients réels a_k ($k = 0, 1, 2, \ldots, n$). Proposons-nous de trouver la valeur de ce polynôme pour $x = \xi$:

$$P(\xi) = a_0\xi^n + a_1\xi^{n-1} + \ldots + a_{n-1}\xi + a_n. \tag{2}$$

Le plus commode pour calculer le nombre $P(\xi)$ est de mettre la formule (2) sous la forme

$$P(\xi) = (\ldots(((a_0\xi + a_1)\xi + a_2)\xi + a_3)\xi + \ldots \\ \ldots + a_{n-1})\xi + a_n).$$

En calculant successivement les nombres

$$\left.\begin{aligned} b_0 &= a_0, \\ b_1 &= a_1 + b_0\xi, \\ b_2 &= a_2 + b_1\xi, \\ b_3 &= a_3 + b_2\xi, \\ &\ldots\ldots\ldots \\ b_n &= a_n + b_{n-1}\xi, \end{aligned}\right\} \tag{3}$$

on aboutit à $b_n = P(\xi)$.

Montrons que les nombres $b_0 = a_0, b_1, \ldots, b_{n-1}$ sont les coefficients du polynôme $Q(x)$ obtenu comme quotient de la division du polynôme donné $P(x)$ par le binôme $x - \xi$. En effet, soit

$$Q(x) = \beta_0 x^{n-1} + \beta_1 x^{n-2} + \ldots + \beta_{n-1} \tag{4}$$

et

$$P(x) = Q(x)(x - \xi) + \beta_n; \tag{5}$$

d'après le théorème de Bézout le reste de la division $\beta_n = P(\xi)$. En vertu des formules (4) et (5)

$$P(x) = (\beta_0 x^{n-1} + \beta_1 x^{n-2} + \ldots + \beta_{n-1})(x - \xi) + \beta_n,$$

ou, en chassant les parenthèses et en effectuant la réduction des termes semblables, on a

$$P(x) = \beta_0 x^n + (\beta_1 - \beta_0\xi) x^{n-1} + (\beta_2 - \beta_1\xi) x^{n-2} + \ldots$$
$$\ldots + (\beta_{n-1} - \beta_{n-2}\xi) x + (\beta_n - \beta_{n-1}\xi).$$

La comparaison des coefficients affectés aux mêmes puissances de la variable x dans le premier et le second membres de cette dernière égalité donne:

$$\begin{aligned} \beta_0 &= a_0, \\ \beta_1 - \beta_0\xi &= a_1, \\ \beta_2 - \beta_1\xi &= a_2, \\ &\ldots\ldots\ldots \\ \beta_{n-1} - \beta_{n-2}\xi &= a_{n-1}, \\ \beta_n - \beta_{n-1}\xi &= a_n. \end{aligned}$$

D'où

$$\begin{aligned} \beta_0 &= a_0 = b_0, \\ \beta_1 &= a_1 + \beta_0\xi = b_1, \\ \beta_2 &= a_2 + \beta_1\xi = b_2, \\ &\ldots\ldots\ldots\ldots \\ \beta_{n-1} &= a_{n-1} + \beta_{n-2}\xi = b_{n-1}, \\ \beta_n &= a_n + \beta_{n-1}\xi = b_n, \end{aligned}$$

ce qu'il fallait démontrer.

Ainsi les formules (3) permettent de déterminer, sans effectuer la division, les coefficients du quotient $Q(x)$ et le reste $P(\xi)$. Les calculs se font pratiquement d'après le schéma suivant, dit *schéma de Hörner*:

$$\begin{array}{r|l} + \begin{array}{ccccc} a_0 & a_1 & a_2 & \ldots & a_n \\ & b_0\xi & b_1\xi & \ldots & b_{n-1}\xi \end{array} & \underline{\xi} \\ \hline \begin{array}{ccccc} b_0 & b_1 & b_2 & \ldots & \underline{b_n = P(\xi)} \end{array} & \end{array}$$

E x e m p l e 1. Calculer la valeur du polynôme

$$P(x) = 3x^3 + 2x^2 - 5x + 7 \text{ avec } x = 3.$$

S o l u t i o n. Composons le schéma de Hörner:

$$\begin{array}{rrrrrl} & 3 & 2 & -5 & 7 & \underline{|3} \\ + & & 9 & 33 & 84 & \\ \hline & 3 & 11 & 28 & \underline{91 = P(3)} & \end{array}$$

R e m a r q u e. En appliquant le schéma de Hörner on peut obtenir les bornes des racines réelles du polynôme considéré $P(x)$.

Fig. 2.

Posons qu'avec $x = \beta$ ($\beta > 0$) tous les coefficients b_i du schéma de Hörner sont non négatifs et que le premier coefficient est positif, c'est-à-dire

$$b_0 = a_0 > 0 \quad \text{et} \quad b_i \geqslant 0 \quad (i = 1, 2, \ldots, n). \tag{6}$$

On peut alors affirmer que toute racine réelle x_k ($k = 1, 2, \ldots, m$; $m \leqslant n$) du polynôme $P(x)$ ne dépasse pas β, c'est-à-dire que $x_k \leqslant \beta$ ($k = 1, 2, \ldots, m$) (fig. 2).

En effet, puisque

$$P(x) = (b_0 x^{n-1} + \ldots + b_{n-1})(x - \beta) + b_n,$$

quel que soit $x > \beta$ et en vertu de la condition (6) on a $P(x) > 0$, ce qui rend évident le fait qu'un nombre quelconque supérieur à β n'est pas une racine du polynôme $P(x)$. On a ainsi une majorante des racines réelles x_k du polynôme.

Pour obtenir une minorante des racines x_k, composons le polynôme

$$(-1)^n P(-x) = a_0 x^n - a_1 x^{n-1} + \ldots + (-1)^n a_n.$$

Cherchons pour ce nouveau polynôme un nombre $x = \alpha$ ($\alpha > 0$) tel que tout coefficient du schéma de Hörner correspondant soit non négatif sauf le premier qui, naturellement, est positif. Alors, pour les racines réelles du polynôme $(-1)^n P(-x)$ égales, évidemment, à $-x_k$ ($k = 1, 2, \ldots, m$), les raisonnements précédents amènent l'inégalité $-x_k \leqslant \alpha$.

Par conséquent, $x_k \geqslant -\alpha$ ($k = 1, 2, \ldots, m$). Ainsi nous avons obtenu la limite inférieure $-\alpha$ des racines réelles du polynôme $P(x)$. On en déduit que toute racine réelle du polynôme $P(x)$ repose dans l'intervalle $[-\alpha, \beta]$.

Exemple 2. Chercher les limites des racines réelles du polynôme

$$P(x) = x^4 - 2x^3 + 3x^2 + 4x - 1.$$

Solution. Calculons la valeur du polynôme $P(x)$ pour $x = 2$, par exemple. En appliquant le schéma de Hörner, on a:

$$\begin{array}{rrrrr|l} 1 & -2 & 3 & 4 & -1 & \\ + & 2 & 0 & 6 & 20 & \underline{2} \\ \hline 1 & 0 & 3 & 10 & \underline{19} & \end{array}$$

Puisque tout coefficient $b_i \geqslant 0$, les racines réelles x_k du polynôme $P(x)$ (si elles existent) vérifient l'inégalité $x_k < 2$. La limite supérieure des racines réelles est ainsi obtenue. Passons à la recherche de la limite inférieure. Composons le nouveau polynôme:

$$Q(x) = (-1)^4 P(-x) = x^4 + 2x^3 + 3x^2 - 4x - 1.$$

Le calcul de sa valeur pour $x = 1$, par exemple, donne:

$$\begin{array}{rrrrr|l} 1 & 2 & 3 & -4 & -1 & \\ + & 1 & 3 & 6 & 2 & \underline{1} \\ \hline 1 & 3 & 6 & 2 & 1 & \end{array}$$

Tout coefficient $b_i > 0$, donc $-x_k < 1$, c'est-à-dire $x_k > -1$. Toutes les racines réelles du polynôme considéré sont comprises donc à l'intérieur de l'intervalle $[-1, 2]$.

§ 2. Schéma de Hörner généralisé

Soit le polynôme

$$P(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \qquad (1)$$

dans lequel, d'après certaines considérations, il faut remplacer la variable x suivant la formule

$$x = y + \xi, \qquad (2)$$

où ξ est un nombre fixé et y une variable nouvelle.

En portant l'expression (2) dans le polynôme (1), après avoir effectué les opérations indiquées et la réduction des termes semblables, on obtient un nouveau polynôme par rapport à la variable y:

$$P(y + \xi) = A_0 y^n + A_1 y^{n-1} + \ldots + A_n. \qquad (3)$$

Le polynôme (3) pouvant être considéré comme le développement de Taylor de la fonction $P(y + \xi)$ suivant les puissances de y, les coefficients A_i $(i = 0, 1, \ldots, n)$ se calculent d'après la formule

$$A_i = \frac{P^{(n-i)}(\xi)}{(n-i)!} \quad (i = 0, 1, \ldots, n).$$

Indiquons un procédé pratique plus commode de la recherche des coefficients A_i ($i = 0, 1, 2, \ldots, n$) à l'aide du schéma de Hörner. Posons d'abord dans l'expression (3) $y = 0$. On a alors $A_n = P(\xi)$.

La division du polynôme (1) par le binôme $x - \xi$ amène

$$P(x) = (x - \xi) P_1(x) + P(\xi), \tag{4}$$

où

$$P_1(x) = b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-1}.$$

Ensuite, si dans l'expression (3) y est remplacé par sa valeur $y = x - \xi$, il vient:

$$P(x) = (x - \xi)[A_0 (x - \xi)^{n-1} + A_1 (x - \xi)^{n-2} + \ldots$$
$$\ldots + A_{n-1}] + P(\xi). \tag{5}$$

La comparaison des formules (4) et (5) conduit à la conclusion que

$$P_1(x) = A_0 (x - \xi)^{n-1} + A_1 (x - \xi)^{n-2} + \ldots + A_{n-1}, \tag{6}$$

d'où

$$A_{n-1} = P_1(\xi). \tag{7}$$

D'une façon analogue, en divisant le polynôme $P_1(x)$ par le binôme $x - \xi$, on peut poser:

$$P_1(x) = (x - \xi) P_2(x) + P_1(\xi), \tag{8}$$

où $P_2(x) = c_0 x^{n-2} + c_1 x^{n-3} + \ldots + c_{n-2}$.

En outre, les formules (6) et (7) entraînent:

$$P_1(x) = (x - \xi)[A_0 (x - \xi)^{n-2} + A_1 (x - \xi)^{n-3} + \ldots$$
$$\ldots + A_{n-2}] + P_1(\xi). \tag{9}$$

La comparaison des formules (8) et (9) fait conclure que

$$P_2(x) = A_0 (x - \xi)^{n-2} + A_1 (x - \xi)^{n-3} + \ldots + A_{n-2}.$$

On en tire $A_{n-2} = P_2(\xi)$.

En poursuivant ce processus nous exprimerons successi ement tous les coefficients A_i ($i = 0, 1, \ldots, n$) par les valeurs des polynômes correspondants $P_0(x) = P(x)$, et $P_1(x), \ldots, P_n(x) = a_0$, avec $x = \xi$:

$$A_n = P(\xi);$$
$$A_{n-1} = P_1(\xi);$$
$$A_{n-2} = P_2(\xi);$$
$$\ldots\ldots\ldots$$
$$A_0 = P_n(\xi).$$

où les polynômes $P_{k+1}(x)$ se construisent, en partant des polynômes $P_k(x)$, d'après la formule

$$P_k(x) = (x - \xi)\, P_{k+1}(x) + P_k(\xi) \qquad (k = 0, 1, \ldots, n).$$

Pour calculer les valeurs des $P(\xi)$, $P_1(\xi)$, $P_2(\xi)$, . . . utilisons le *schéma de Hörner généralisé*:

$$\begin{array}{ccccccc|l}
a_0 & a_1 & a_2 & a_3 & \ldots & a_{n-1} & a_n & \xi \\
 & b_0\xi & b_1\xi & b_2\xi & \ldots & b_{n-2}\xi & b_{n-1}\xi & \\
\hline
b_0 & b_1 & b_2 & b_3 & \ldots & b_{n-1} & \underline{b_n = P(\xi)} & \\
 & c_0\xi & c_1\xi & \ldots & & c_{n-2}\xi & & \\
\hline
c_0 & c_1 & c_2 & \ldots & & \underline{c_{n-1} = P_1(\xi)} & & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & &
\end{array}$$

E x e m p l e. Pour éliminer du polynôme

$$P(x) = x^4 - 8x^3 + 5x^2 + 2x - 7$$

le terme contenant la variable à la troisième puissance, on pose $x = y + 2$.

Trouver le polynôme transformé.

S o l u t i o n. Appliquons le schéma de Hörner généralisé:

1	−8	5	2	−7	$\xi = 2$
	2	−12	−14	−24	
1	−6	−7	−12	<u>−31</u>	
	2	−8	−30		
1	−4	−15	<u>−42</u>		
	2	−4			
1	−2	<u>−19</u>			
	2				
1	<u>0</u>				
<u>1</u>					

Par conséquent,

$$P(y + 2) = y^4 - 19y^2 - 42y - 31.$$

§ 3. Calcul des fractions rationnelles

Toute *fraction rationnelle* $R(x)$ peut être mise sous la forme d'un rapport de deux polynômes

$$R(x) = \frac{P(x)}{Q(x)}, \tag{1}$$

où

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n,$$
$$Q(x) = b_0x^m + b_1x^{m-1} + \ldots + b_m.$$

Supposons qu'on demande de déterminer la valeur de $R(x)$ en un point $x = \xi$:

$$R(\xi) = \frac{P(\xi)}{Q(\xi)}. \tag{2}$$

Le numérateur $P(\xi)$ et le dénominateur $Q(\xi)$ de la fraction (2) peuvent s'obtenir en recourant au schéma de Hörner. Il en résulte une méthode simple de calcul du nombre $R(\xi)$.

Une autre méthode de calcul consiste à transformer la fonction $R(x)$ en fraction continue. A cette fin on peut appliquer la méthode décrite au § 3 du chapitre II.

§ 4. Approximation des sommes des séries numériques

D é f i n i t i o n. La série numérique

$$a_1 + a_2 + \ldots + a_n + \ldots \tag{1}$$

s'appelle *convergente* si la suite de ses sommes partielles a une limite

$$S = \lim_{n\to\infty} S_n, \tag{2}$$

où

$$S_n = a_1 + a_2 + \ldots + a_n.$$

Le nombre S s'appelle *somme de la série.*

Ainsi, la convergence de la série (1) est équivalente à la convergence de la suite de ses sommes partielles. D'après le *critère de Cauchy* [1] cette suite converge si et seulement si pour tout $\varepsilon > 0$ il existe un $N = N(\varepsilon)$ tel que

$$|S_{n+p} - S_n| < \varepsilon$$

avec $n > N$ et $p > 0$ arbitraire.

La formule (2) entraîne

$$S = S_n + R_n, \tag{3}$$

où R_n est le *reste de la série* et $R_n \to 0$ lorsque $n \to \infty$.

Pour obtenir la somme S de la série convergente (1) avec la précision imposée ε, le nombre de termes n doit être suffisamment grand pour vérifier l'inégalité

$$|R_n| < \varepsilon. \tag{4}$$

On peut alors poser approximativement que la somme partielle S_n est la somme précise S de la série (1).

Constatons que pratiquement la détermination des termes a_1, a_2, ... est également approximative. De plus, la somme S_n est d'habitude arrondie au nombre imposé de chiffres. Pour tenir compte de toutes ces erreurs et assurer la précision nécessaire, on peut appliquer la procédure suivante: choisissons dans le cas général trois

nombres positifs ε_1, ε_2, ε_3 tels que

$$\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = \varepsilon.$$

Prenons le nombre n de termes de la série suffisamment grand pour que l'*erreur de troncature* $|R_n|$ vérifie l'inégalité

$$|R_n| \leqslant \varepsilon_1. \tag{5}$$

Calculons chacun des termes a_k ($k = 1, 2, \ldots, n$) avec une borne d'erreur absolue ne dépassant pas $\frac{\varepsilon_2}{n}$, et soit $\overline{a}_k$ ($k = 1, 2, \ldots, n$) les valeurs approchées correspondantes des termes de la série (1), c'est-à-dire

$$|\overline{a}_k - a_k| \leqslant \frac{\varepsilon_2}{n}.$$

Alors l'*erreur générée* de la sommation

$$\overline{S}_n = \sum_{k=1}^{n} \overline{a}_k$$

vérifie l'inégalité

$$|S_n - \overline{S}_n| \leqslant \varepsilon_2. \tag{6}$$

Enfin, arrondissons le résultat approché obtenu $\overline{S}_n$ à un nombre plus simple $\overline{\overline{S}}_n$ de façon que l'*erreur d'arrondi* soit

$$|\overline{S}_n - \overline{\overline{S}}_n| \leqslant \varepsilon_2. \tag{7}$$

Dans ce cas le nombre $\overline{\overline{S}}_n$ est une valeur approchée de la somme S de la série (1) à ε près. En effet, les inégalités (5)-(7) amènent

$$|S - \overline{\overline{S}}_n| \leqslant |S - S_n| + |S_n - \overline{S}_n| + |\overline{S}_n - \overline{\overline{S}}_n| \leqslant \varepsilon_1 + \varepsilon_2 + \varepsilon_3 = \varepsilon.$$

Le nombre ε doit être partitionné en termes positifs ε_1, ε_2 et ε_3 en tenant compte du volume de travail nécessaire pour obtenir le résultat cherché. Si $\varepsilon = 10^{-m}$ et le résultat doit compter m décimales exactes, on adopte le plus souvent :

$$\varepsilon_1 = \frac{\varepsilon}{4}, \quad \varepsilon_2 = \frac{\varepsilon}{4}, \quad \varepsilon_3 = \frac{\varepsilon}{2}.$$

Si l'on ne procède pas à l'arrondissement final, on pose dans les cas courants :

$$\varepsilon_1 = \frac{\varepsilon}{2}, \quad \varepsilon_2 = \frac{\varepsilon}{2}, \quad \varepsilon_3 = 0.$$

La tâche se complique lorsqu'il faut déterminer la somme de la série avec m décimales exactes au sens strict. L'interprétation géométrique de ce fait est qu'il faut chercher un élément de

l'ensemble des $\frac{k}{10^m}$ (k est un entier) tel qu'il soit le plus proche du nombre S (fig. 3).

Fig. 3.

Soit, pour fixer les idées, la somme S positive et

$$\tilde{S} = p_0 + \frac{p_1}{10} + \ldots + \frac{p_m}{10^m} + \ldots + \frac{p_n}{10^n}$$

(p_k sont des entiers non négatifs, $n \geqslant m$) une approximation rationnelle telle que

$$|S - \tilde{S}| \leqslant \frac{1}{10^{m+1}}.$$

Admettons que

$$p_{m+1} \neq 4 \quad \text{et} \quad p_{m+1} \neq 5.$$

Alors, en arrondissant le nombre $\tilde{S}$ on aboutit au résultat recherché:

$$\sigma = p_0 + \frac{p_1}{10} + \ldots + \frac{p_m}{10^m}, \quad \text{si } p_{m+1} \leqslant 3, \tag{8}$$

ou

$$\sigma = p_0 + \frac{p_1}{10} + \ldots + \frac{p_m + 1}{10^m}, \quad \text{si } p_{m+1} \geqslant 6. \tag{8'}$$

En effet, dans le premier cas, l'arrondissement par défaut conduit à

$$0 \leqslant \tilde{S} - \sigma = \frac{p_{m+1}}{10^{m+1}} + \frac{p_{m+2}}{10^{m+2}} + \ldots + \frac{p_n}{10^n} \leqslant$$
$$\leqslant \frac{3}{10^{m+1}} + \frac{9}{10^{m+2}} + \ldots + \frac{9}{10^n} < \frac{4}{10^{m+1}}.$$

Dans le deuxième cas, l'arrondissement par excès donne:

$$0 \leqslant \sigma - \tilde{S} = \frac{1}{10^m} - \frac{p_{m+1}}{10^{m+1}} - \ldots - \frac{p_n}{10^n} \leqslant \frac{1}{10^m} - \frac{6}{10^{m+1}} = \frac{4}{10^{m+1}}.$$

C'est pourquoi dans les deux cas on a:

$$|\tilde{S} - \sigma| \leqslant \frac{4}{10^{m+1}}$$

et, par conséquent,

$$|S - \sigma| \leqslant |S - \tilde{S}| + |\tilde{S} - \sigma| \leqslant \frac{1}{10^{m+1}} + \frac{4}{10^{m+1}} = \frac{1}{2} \cdot 10^{-m}.$$

Ainsi,

$$S = \sigma \pm \frac{1}{2} \cdot 10^{-m}.$$

Si $p_{m+1} = 4$ ou $p_{m+1} = 5$, il faut améliorer la précision des calculs de la somme approchée S en faisant appel au rang décimal suivant.

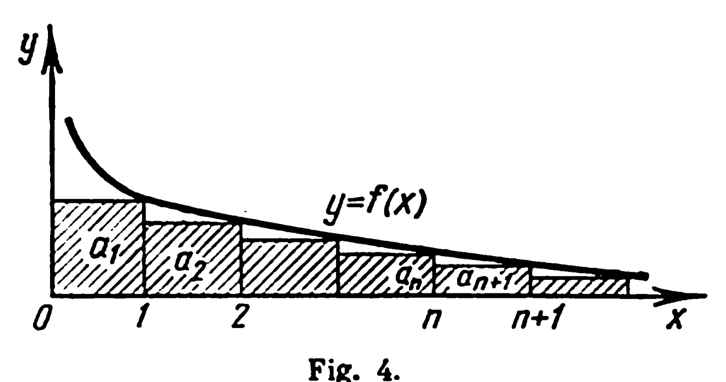


Fig. 4.

Dans le cas particulier, si $p_{m+1} = 4$ et si l'on sait que

$$S < \tilde{S},$$

σ (8) est une valeur approchée de la somme S à $\frac{1}{2} \cdot 10^{-m}$ près (par défaut).

D'une façon analogue, si $p_{m+1} = 5$ et

$$S > \tilde{S},$$

σ (8′) est une valeur approchée de la somme S à $\frac{1}{2} \cdot 10^{-m}$ près (par excès).

Pour évaluer le reste de la série (1)

$$R_n = a_{n+1} + a_{n+2} + \ldots$$

il est utile de faire appel aux théorèmes suivants que nous donnons sans démonstration [1].

Théorème 1. *Si les termes de la série* (1) *sont les valeurs correspondantes d'une fonction* $f(x)$ *monotone décroissante positive, c'est-à-dire si*

$$a_n = f(n) \quad (n = 1, 2, \ldots), \tag{9}$$

alors (fig. 4)

$$\int_{n+1}^{\infty} f(x)\, dx < R_n < \int_{n}^{\infty} f(x)\, dx.$$

Théorème 2. *Si la série* (1) *est alternée:*

$$a_1 > 0,\ a_2 < 0,\ a_3 > 0,\ \ldots$$

et si les modules de ses termes forment une suite monotone décroissante, on a

$$|R_n| \leqslant |a_{n+1}|$$

et

$$\operatorname{sgn} R_n = \operatorname{sgn} a_{n+1} \text{ *}.$$

E x e m p l e. Trouver la somme de la série

$$S = \frac{1}{1^3} + \frac{1}{2^3} + \frac{1}{3^3} + \ldots + \frac{1}{n^3} + \ldots \tag{10}$$

à 0,001 près.

S o l u t i o n. Adoptons comme erreur de troncature

$$\varepsilon_1 = \frac{1}{4} \cdot 10^{-3} = \frac{1}{4000}.$$

Les termes de la série (10) sont les valeurs correspondantes de la fonction décroissante monotone

$$f(x) = \frac{1}{x^3}.$$

C'est pourquoi pour le n-ième reste de la série

$$R_n = \sum_{k=n+1}^{\infty} \frac{1}{k^3}$$

on a l'estimation

$$R_n \leqslant \int_n^{\infty} \frac{dx}{x^3} = \frac{1}{2n^2}.$$

La résolution de l'inégalité

$$\frac{1}{2n^2} \leqslant \frac{1}{4000}$$

conduit à :

$$n \geqslant \sqrt{2\,000} \approx 44{,}7.$$

Adoptons $n = 45$.

Choisissons comme borne d'erreur de la sommation

$$\varepsilon_2 = \frac{1}{4} \cdot 10^{-3};$$

il en résulte que la borne d'erreur absolue admissible des termes de la somme partielle S_{45} de la série (10) est

$$\frac{\varepsilon_2}{n} \leqslant \frac{\frac{1}{4} \cdot 10^{-3}}{45} = \frac{5}{9} \cdot 10^{-5}.$$

* sgn R_n désigne le *signe* du nombre R_n, c'est-à-dire sgn $R_n = +1$ si $R_n > 0$, sgn $R_n = -1$ si $R_n < 0$, sgn $R_n = 0$ si $R_n = 0$.

Posons

$$\frac{\varepsilon_2}{n} = \frac{1}{2} \cdot 10^{-5},$$

c'est-à-dire calculons les termes de la série (10) avec cinq décimales exactes au sens strict. Ci-dessous figurent les valeurs correspondantes des termes et les résultats de la sommation partielle

1,00000	0,00024	0,00003
0,12500	0,00020	0,00003
0,03704	0,00017	0,00003
0,01562	0,00014	0,00003
0,00800	0,00012	0,00002
0,00463	0,00011	0,00002
0,00292	0,00009	0,00002
0,00195	0,00008	0,00002
0,00137	0,00007	0,00002
0,00100	0,00006	0,00002
0,00075	0,00006	0,00001
0,00058	0,00005	0,00001
0,00046	0,00004	0,00001
0,00036	0,00004	0,00001
0,00030	0,00004	0,00001
1,19998	0,00151	0,00029

Par conséquent,

$$S_{45} = 1,19998 + 0,00151 + 0,00029 = 1,20178.$$

En arrondissant cette valeur à des millièmes, on obtient la valeur approchée de la somme

$$S \approx 1,202.$$

L'erreur d'arrondi étant

$$\varepsilon_3 = 0,00022 < \frac{1}{4} \cdot 10^{-3},$$

l'erreur globale du résultat obtenu ne dépasse pas

$$\varepsilon < \frac{1}{4} \cdot 10^{-3} + \frac{1}{4} \cdot 10^{-3} + \frac{1}{4} \cdot 10^{-3} < \frac{3}{4} \cdot 10^{-3}.$$

Ainsi,

$$S = 1,202 \pm 0,001.$$

L'estimation sera plus précise si l'on tient compte des signes des erreurs d'arrondi. Pour comparer, voici la valeur de la somme S à $\frac{1}{2} \cdot 10^{-6}$ près [2]:

$$S = 1,202057.$$

R e m a r q u e. La recherche de l'erreur globale étant une opération très délicate, dans la pratique la précision imposée $\varepsilon = 10^{-m}$ s'obtient en effectuant tous les calculs intermédiaires avec un ou deux chiffres de réserve. Dans ces conditions on suppose sans trop de rigueur que les erreurs admises n'interviennent pas dans les décimales de rang m du résultat à obtenir.

Remarquons que pour résoudre cet exemple il faut chercher la somme d'un nombre de termes relativement grand. Dans la pratique, on s'efforce de transformer la série considérée de façon à obtenir le résultat cherché avec un petit nombre de termes. Les transformations de ce type s'appellent *amélioration de la convergence d'une série* et dans de nombreux cas elles permettent de réduire nettement la durée du calcul. Cette question fait l'objet du chapitre VI.

§ 5. Fonctions analytiques

Une fonction réelle $f(x)$ s'appelle fonction analytique au point ξ si dans un certain voisinage $|x - \xi| < R$ de ce point la fonction se développe en une série entière (*serie de Taylor*)

$$f(x) = f(\xi) + f'(\xi)(x-\xi) + \frac{f''(\xi)}{2!}(x-\xi)^2 + \ldots \\ \ldots + \frac{f^{(n)}(\xi)}{n!}(x-\xi)^n + \ldots \quad (1)$$

Avec $\xi = 0$ on obtient la *série de Maclaurin*

$$f(x) = f(0) + f'(0)\,x + \frac{f''(0)}{2!}x^2 + \ldots + \frac{f^{(n)}(0)}{n!}x^n + \ldots \quad (2)$$

La différence

$$R_n(x) = f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(\xi)}{k!}(x-\xi)^k$$

s'appelle *reste* et constitue l'erreur produite en remplaçant la fonction $f(x)$ par le *polynôme de Taylor*

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(\xi)}{k!}(x-\xi)^k.$$

On sait que [1]

$$R_n(x) = \frac{f^{(n+1)}(\xi + \theta(x-\xi))}{(n+1)!}(x-\xi)^{n+1}, \quad (3)$$

où $0 < \theta < 1$. Pour la série de Maclaurin (2) on a en particulier [1]:

$$R_n(x) = \frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1}, \quad (4)$$

où $0 < \theta < 1$. Il existe également d'autres formes de restes.

Le développement d'une fonction en série de Taylor est dans plusieurs cas un moyen commode pour calculer les valeurs de cette fonction.

Si l'on connaît $f(\xi)$ et s'il faut trouver la valeur de $f(\xi+h)$, où h est une « petite correction », il est commode d'écrire la formule (1) sous la forme

$$f(\xi+h)=f(\xi)+f'(\xi)h+\frac{f''(\xi)}{2!}h^2+\ldots+\frac{f^{(n)}(\xi)}{n!}h^n+R_n(h), \tag{5}$$

où

$$R_n(h)=\frac{f^{(n+1)}(\xi+\theta h)}{(n+1)!}h^{n+1} \quad (0<\theta<1).$$

E x e m p l e. Trouver la valeur approchée de $\sqrt{10}$.

S o l u t i o n. On a

$$\sqrt{10}=\sqrt{3^2+1}=3\left(1+\frac{1}{9}\right)^{\frac{1}{2}}. \tag{6}$$

En posant

$$f(x)=(1+x)^{\frac{1}{2}},$$

on obtient successivement:

$$f'(x)=\frac{1}{2}(1+x)^{-\frac{1}{2}},$$

$$f''(x)=-\frac{1}{4}(1+x)^{-\frac{3}{2}},$$

$$f'''(x)=\frac{3}{8}(1+x)^{-\frac{5}{2}},$$

$$f^{\mathrm{IV}}(x)=-\frac{15}{16}(1+x)^{-\frac{7}{2}}.$$

D'où, en posant $\xi=0$, $h=\frac{1}{9}$ et en tenant compte du fait que

$$f(0)=1,\quad f'(0)=\frac{1}{2},\quad f''(0)=-\frac{1}{4},\quad f'''(0)=\frac{3}{8},$$

on a en vertu de la formule (5):

$$\left(1+\frac{1}{9}\right)^{\frac{1}{2}}=1+\frac{1}{2}\cdot\frac{1}{9}-\frac{1}{8}\cdot\frac{1}{81}+\frac{1}{16}\cdot\frac{1}{729}+R_3=$$
$$=1+0,05556-0,00154+0,00009+R_3=1,05411+R_3, \tag{7}$$

où

$$R_3 = -\frac{1}{24}\cdot\frac{15}{16}\cdot\left(1+\frac{\theta}{9}\right)^{-\frac{7}{2}}\cdot\frac{1}{6561} =$$

$$= -\frac{10}{1\,680\,616}\cdot\left(1+\frac{\theta}{9}\right)^{-\frac{7}{2}} \quad (0<\theta<1).$$

Il est clair que

$$|R_3| < \frac{10}{1\,680\,616} < 6\cdot 10^{-6}.$$

Les formules (6) et (7) entraînent:

$$\sqrt{10} = 3{,}16233 + E, \tag{8}$$

où

$$|E| < 3\cdot\frac{1}{2}\cdot 10^{-5} + 3\cdot 6\cdot 10^{-6} = 3{,}3\cdot 10^{-5}.$$

En arrondissant la valeur obtenue à quatre décimales, on a finalement:

$$\sqrt{10} = 3{,}1623 \pm 6\cdot 10^{-5}.$$

A titre de comparaison voici la valeur tabulaire

$$\sqrt{10} = 3{,}1622777\ \ldots$$

§ 6. Fonction exponentielle

Le développement [1] de la fonction exponentielle e^x s'écrit

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots, \tag{1}$$

l'intervalle de convergence étant $-\infty < x < +\infty$. Le reste de la série (1) est de la forme

$$R_n(x) = \frac{e^{\theta x}}{(n+1)!}x^{n+1} \quad (0<\theta<1). \tag{2}$$

Lorsque les modules des valeurs de x sont grands, la série (1) est peu utile pour le calcul. Pour cette raison, on opère en général de la façon suivante: soit

$$x = E(x) + q,$$

où $E(x)$ est la partie entière du nombre x et $0 \leqslant q < 1$ sa partie fractionnaire. On a:

$$e^x = e^{E(x)}\cdot e^q. \tag{3}$$

Le premier facteur du produit (3) peut être établi par multiplication:

$$e^{E(x)} = \overbrace{ee\ \ldots\ e}^{E(x)\ \text{fois}}, \quad \text{si } E(x) \geqslant 0,$$

ou

$$e^{E(x)} = \overbrace{\frac{1}{e} \cdot \frac{1}{e} \cdots \frac{1}{e}}^{-E(x)\ \text{fois}}, \quad \text{si } E(x) < 0,$$

où

$$e = 2{,}71828\,18284\,59045\ldots$$

et

$$\frac{1}{e} = 0{,}36787\,94411\,71442\ldots$$

De plus, pour assurer la précision imposée, il faut prendre e ou $\frac{1}{e}$ avec un nombre de décimales suffisamment grand (actuellement le nombre e est calculé avec plus de 250 décimales).

Quant au deuxième facteur e^q du produit (3), on le calcule à l'aide du développement ci-dessus:

$$e^q = \sum_{n=0}^{\infty} \frac{q^n}{n!}, \tag{4}$$

qui avec $0 \leqslant q < 1$ forme une série rapidement convergente du fait que pour le reste $R_n(q)$ la formule (2) donne l'estimation

$$0 \leqslant R_n(q) < \frac{3}{(n+1)!} q^{n+1}.$$

Déduisons une formule plus précise pour l'estimation du reste $R_n(q)$ avec $0 < q < 1$. On a:

$$R_n(q) = \frac{q^{n+1}}{(n+1)!} + \frac{q^{n+2}}{(n+2)!} + \frac{q^{n+3}}{(n+3)!} + \ldots =$$
$$= \frac{q^{n+1}}{(n+1)!}\left[1 + \frac{q}{n+2} + \frac{q^2}{(n+2)(n+3)} + \ldots\right] <$$
$$< \frac{q^{n+1}}{(n+1)!}\left[1 + \frac{q}{n+2} + \left(\frac{q}{n+2}\right)^2 + \ldots\right].$$

On en tire, en sommant la progression géométrique entre crochets:

$$R_n(q) < \frac{q^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{q}{n+2}}; \tag{5}$$

ou avec $0 < q < 1$ et retenant que

$$\frac{n+2}{n+1} < \frac{n+1}{n},$$

on a finalement:

$$0 < R_n(q) < \frac{q^{n+1}}{n!\,n},$$

c'est-à-dire

$$0 < R_n(q) < u_n \cdot \frac{q}{n}, \tag{6}$$

où $u_n = \frac{q^n}{n!}$ est le dernier terme conservé.

Si l'erreur de troncature ε est imposée, le nombre nécessaire de termes n peut être déterminé par triage en résolvant l'inéquation

$$\frac{q^{n+1}}{n!\,n} < \varepsilon.$$

Le calcul approché de e^x d'après la formule (1) pour des x petits devient commode si l'on utilise le schéma

$$e^x = u_0 + u_1 + u_2 + \ldots + u_n + R_n(x), \tag{7}$$

où

$$u_0 = 1, \quad u_k = \frac{xu_{k-1}}{k} \quad (k = 1, 2, \ldots, n). \tag{8}$$

Le calcul sur ordinateurs devient commode si l'on applique le schéma suivant

$$u_k = \frac{x}{k} u_{k-1},$$

$$s_k = s_{k-1} + u_k \quad (k = 0, 1, 2, \ldots, n),$$

où $u_0 = 1$, $s_{-1} = 0$, $s_0 = 1$. Le nombre $s_n = \sum_{k=0}^{n} \frac{x^k}{k!}$ donne approximativement le résultat cherché de e^x.

Si ε est l'erreur de troncature admissible et $n \geqslant 2|x| > 0$, le processus de sommation doit être arrêté dès que l'inégalité

$$|R_n(x)| \leqslant R_n(|x|) < \frac{|x|^{n+1}}{(n+1)!} \cdot \frac{1}{1 - \frac{|x|}{n+2}} < \frac{2|x|^{n+1}}{(n+1)!} = \frac{2|x|}{n+1} \cdot \frac{|x|^n}{n!} < |u_n| \leqslant \varepsilon$$

est vérifiée, c'est-à-dire si

$$|u_n(x)| \leqslant \varepsilon. \tag{9}$$

Autrement dit, le processus de sommation s arrête si le dernier terme élaboré u_n ne dépasse pas ε en module. Alors

$$|R_n(x)| < |u_n|.$$

Pour calculer l'erreur globale, il faut faire appel au schéma général (§ 4).

E x e m p l e 1. Chercher $\sqrt{e}$ à 10^{-5} près.

S o l u t i o n. Adoptons l'erreur de troncature

$$\varepsilon_1 = \frac{1}{4}\cdot 10^{-5} = 2{,}5\cdot 10^{-6}.$$

Le nombre de termes de la somme (7) étant alors, d'après une estimation grossière, de l'ordre de 10, calculons les termes à $\frac{1}{2}\cdot 10^{-7}$ près, c'est-à-dire avec deux décimales de réserve.

En posant

$$u_0 = 1,\quad u_k = \frac{u_{k-1}}{2k} \quad (k = 1,\ 2,\ \ldots),$$

on a successivement:

$$\left.\begin{aligned} u_0 &= \quad 1 \\ u_1 &= \frac{1}{2} = 0{,}5000000 \\ u_2 &= \frac{u_1}{4} = 0{,}1250000 \\ u_3 &= \frac{u_2}{6} = 0{,}0208333 \\ u_4 &= \frac{u_3}{8} = 0{,}0026042 \\ u_5 &= \frac{u_4}{10} = 0{,}0002604 \\ u_6 &= \frac{u_5}{12} = 0{,}0000217 \\ u_7 &= \frac{u_6}{14} = 0{,}0000016 \\ \hline & \quad 1{,}6487212 \end{aligned}\right\}$$

En arrondissant la somme à cinq décimales, on obtient:

$$\sqrt{e} = 1{,}64872, \tag{10}$$

avec une erreur globale

$$\varepsilon < 1{,}6\cdot 10^{-6} + 5\cdot\frac{1}{2}\cdot 10^{-7} + 1{,}2\cdot 10^{-6} = 3{,}05\cdot 10^{-6} < \frac{1}{2}\cdot 10^{-5},$$

c'est-à-dire tous les chiffres du résultat (10) sont exacts au sens strict.

Pour calculer e^x on peut utiliser également le développement en fraction continue [4]

$$e^x = \left[0;\ \frac{1}{1},\ \frac{-2x}{2+x},\ \frac{x^2}{6},\ \frac{x^2}{10},\ \ldots,\ \frac{x^3}{4n+2},\ \ldots\right], \tag{11}$$

qui converge pour toute valeur de x (réelle ou complexe).

E x e m p l e 2. Chercher $\sqrt{e}$ en appliquant la formule (11)

S o l u t i o n. En posant dans la formule (11) $x = \frac{1}{2}$, composons le tableau des fractions correspondantes pour la fraction continue respective.

k	-1	0	1	2	3	4	5
b_k		0	1	-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
a_k	1	1	1	$\frac{5}{2}$	6	10	14
P_k	1	0	1	$\frac{5}{2}$	$\frac{61}{4}$	$\frac{1225}{8}$	$\frac{34\,361}{16}$
Q_k	0	1	1	$\frac{3}{2}$	$\frac{37}{4}$	$\frac{743}{8}$	$\frac{20\,841}{16}$

En s'arrêtant à la cinquième fraction correspondante, on a:

$$\sqrt{e} \approx \frac{P_5}{Q_5} = \frac{34361}{16} : \frac{20841}{16} = \frac{34361}{20841} = 1{,}648721$$

à $\frac{1}{2} \cdot 10^{-6}$ près.

Pour calculer les valeurs de la fonction exponentielle générale a^x $(a > 0)$ on peut recourir à la formule

$$a^x = 1 + \ln a \cdot x + \frac{\ln^2 a}{2!} x^2 + \frac{\ln^3 a}{3!} x^3 + \ldots$$

§ 7. Fonctions logarithmiques

Les logarithmes népériens des nombres proches de l'unité donnent lieu au développement [1]

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$$
$$\ldots + (-1)^{n-1} \frac{x^n}{n} + \ldots \quad (-1 < x \leqslant 1). \quad (1)$$

La formule (1) est peu commode pour le calcul car la marge des nombres $0 < 1 + x \leqslant 2$ n'est pas grande et, par ailleurs, avec $|x|$ proche de l'unité la série (1) converge lentement.

Introduisons une formule plus commode pour le calcul des logarithmes népériens des nombres. En remplaçant dans la formule (1) x par $-x$, on a :

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots - \frac{x^n}{n} - \ldots \qquad (2)$$

En retranchant membre à membre la formule (2) de la formule (1) on aboutit à

$$\ln\frac{1-x}{1+x} = -2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots\right).$$

En posant

$$\frac{1-x}{1+x} = z,$$

on obtient :

$$x = \frac{1-z}{1+z}$$

et, par conséquent,

$$\ln z = -2\left[\frac{1-z}{1+z} + \frac{1}{3}\left(\frac{1-z}{1+z}\right)^3 + \frac{1}{5}\left(\frac{1-z}{1+z}\right)^5 + \ldots\right] \qquad (3)$$

avec $0 < z < +\infty$.

Soit x un nombre positif. Mettons-le sous la forme

$$x = 2^m \cdot z,$$

où m est un entier et $\frac{1}{2} \leqslant z < 1$. Alors, en posant

$$\frac{1-z}{1+z} = \xi,$$

où

$$0 < \xi \leqslant \frac{1-\frac{1}{2}}{1+\frac{1}{2}} = \frac{1}{3},$$

et en appliquant la formule (3), on a :

$$\ln x = \ln 2^m z = m \ln 2 + \ln z =$$
$$= m \ln 2 - 2\left(\xi + \frac{\xi^3}{3} + \ldots + \frac{\xi^{2n-1}}{2n-1}\right) - R_n,$$

où

$$R_n = 2\left(\frac{\xi^{2n+1}}{2n+1} + \frac{\xi^{2n+3}}{2n+3} + \frac{\xi^{2n+5}}{2n+5} + \ldots\right) <$$
$$< 2 \cdot \frac{\xi^{2n+1}}{2n+1}(1 + \xi^2 + \xi^4 + \ldots) < \frac{2}{1-\xi^2} \cdot \frac{\xi^{2n+1}}{2n+1}.$$

Avec $0 < \xi \leqslant \frac{1}{3}$, on obtient :

$$\frac{2}{1-\xi^2} \leqslant \frac{9}{4},$$

et c'est pourquoi

$$0 < R_n < \frac{9}{4} \cdot \frac{\xi^{2n+1}}{2n+1} \qquad (4)$$

ou, plus grossièrement,

$$0 < R_n < \frac{1}{4(2n+1)} \cdot \left(\frac{1}{3}\right)^{2n-1}.$$

En introduisant la notation

$$u_k = \frac{\xi^{2k-1}}{2k-1} \qquad (k = 1, 2, \ldots),$$

il vient:

$$\ln x = m \ln 2 - 2(u_1 + u_2 + \ldots + u_n) - R_n, \qquad (5)$$

où

$$\ln 2 = 0{,}69314718\ldots$$

La procédure de sommation s'arrête dès que

$$u_n < 4\varepsilon,$$

où ε est l'erreur de troncature admissible, du fait qu'en vertu de la formule (4),

$$R_n < \frac{9}{4}\xi^2 \cdot \frac{\xi^{2n-1}}{2n-1} \leqslant \frac{1}{4} u_n < \varepsilon.$$

Pour évaluer la borne d'erreur de la somme $\sum_{k=1}^{n} u_k$ on peut se donner un certain nombre de chiffres des termes de la somme et établir approximativement d'après la formule (4) le nombre de termes n.

Exemple. Chercher ln 3 à 10^{-5} près.

Solution. Réalisons le calcul avec deux décimales de réserve. Posons

$$3 = 2^2 \cdot \frac{3}{4} = 2^2 \cdot 0{,}75.$$

Il en résulte que $z = 0{,}75$ et

$$\xi = \frac{1-z}{1+z} = \frac{0{,}25}{1{,}75} = \frac{1}{7} = 0{,}1428571.$$

On a:

$$\left.\begin{aligned} u_1 &= \ \xi = 0{,}1428571 \\ u_2 &= \frac{\xi^3}{3} = 0{,}0009718 \\ u_3 &= \frac{\xi^5}{5} = 0{,}0000119 \\ u_4 &= \frac{\xi^7}{7} = 0{,}0000002 \end{aligned}\right\}$$

$$\overline{\quad 0{,}1438410 \quad}$$

L'application de la formule (5) amène:

$$\ln 3 = 2 \cdot 0{,}69314718 - 2 \cdot 0{,}1438410 = 1{,}09861.$$

Remarque. Les logarithmes népériens des nombres peuvent être également calculés d'après la représentation

$$x = e^p z,$$

où p est un entier et $\frac{1}{e} < z \leqslant 1$ (cf. [5]).

Pour calculer les logarithmes décimaux on utilise la formule suivante

$$\lg x = M \ln x,$$

où

$$M = \lg e = 0{,}43429\ 44819\ 03252\ldots$$

§ 8. Fonctions trigonométriques

A. Calcul des valeurs de sinus et de cosinus

Les formules de réduction permettent d'inclure l'argument x dans l'intervalle $0 \leqslant x \leqslant \frac{\pi}{2}$. Si $0 \leqslant x \leqslant \frac{\pi}{4}$, on a:

$$\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}, \tag{1}$$

mais si $\frac{\pi}{4} \leqslant x \leqslant \frac{\pi}{2}$, on pose

$$\sin x = \cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, \tag{2}$$

où $z = \frac{\pi}{2} - x$ et $0 \leqslant z \leqslant \frac{\pi}{4}$.

Pour calculer la somme de la série (1), il est commode de mettre en œuvre la procédure de sommation

$$\sin x = u_1 + u_2 + \ldots + u_n + R_n, \tag{3}$$

où les termes u_k ($k = 1, 2, \ldots, n$) s'obtiennent successivement par récurrence

$$u_1 = x, \quad u_{k+1} = -\frac{x^2}{2k(2k+1)} u_k \quad (k = 1, 2, \ldots, n-1).$$

(1) étant une série alternée aux termes décroissant en module, le reste R_n vérifie l'estimation

$$|R_n| \leqslant \frac{x^{2n+1}}{(2n+1)!} = |u_{n+1}|$$

et

$$\operatorname{sgn} R_n = \operatorname{sgn} u_{n+1}.$$

La procédure de sommation peut donc être arrêtée dès que

$$|u_n| \leqslant \varepsilon,$$

où ε est l'erreur de troncature imposée.

D'une façon analogue,

$$\cos z = v_1 + v_2 + \ldots + v_n + R_n,$$

où

$$v_1 = 1, \; v_{k+1} = -\frac{x^2}{(2k-1)\,2k} v_k \quad (k = 1, 2, \ldots, n-1)$$

et

$$|R_n| \leqslant \frac{z^{2n}}{(2n)!} = |v_{n+1}|, \; \operatorname{sgn} R_n = \operatorname{sgn} v_{n+1}.$$

Exemple. Chercher sin 20°30′ à 10^{-5} près.

Solution. On a:

$$x = \operatorname{arc} 20°30' = \frac{\pi}{9} + \frac{\pi}{360} = 0{,}349066 + 0{,}008727 = 0{,}357793.$$

En appliquant la formule (3) on obtient:

$$\left.\begin{aligned} u_1 &= \;\; x \;\; = \;\; 0{,}357793 \\ u_2 &= \frac{x^2u_1}{2\cdot 3} = -0{,}007634 \\ u_3 &= \frac{x^2u_2}{4\cdot 5} = +0{,}000049 \\ u_4 &= \frac{x^2u_3}{6\cdot 7} = -0{,}000000 \end{aligned}\right\}$$

$$\overline{0{,}350208},$$

d'où

$$\sin 20°30' = 0{,}35021.$$

D'une façon analogue on détermine les valeurs du $\cos x$.

B. Calcul de la tangente

On peut considérer que $0 \leqslant x \leqslant \frac{\pi}{4}$. Avec $|x| < \frac{\pi}{2}$, $\operatorname{tg} x$ vérifie le développement [6]

$$\operatorname{tg} x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \ldots$$

Les coefficients du développement s'expriment par les *nombres de Bernoulli* (cf. chapitre XVI, § 12).

Le calcul de la valeur de la tangente est très simple si l'on utilise les fractions continues. En posant

$$\operatorname{tg} x = \frac{x}{y},$$

on a d'après la formule de Lambert (chapitre II, § 6)

$$y = \left[1;\ \frac{-x^2}{3},\ \frac{-x^2}{5},\ \ldots,\ \frac{-x^2}{2n+1},\ \ldots\right].$$

Pour calculer y à 10^{-10} près il suffit de poser $n = 7$. Il vient

$$y = 1 - x^2 : (3 - x^2 : (5 - x^2 : (7 - x^2 : (9 - \\ - x^2 : (11 - x^2 : (13 - x^2 : 15)))))). \qquad (4)$$

Le calcul de y se fait en général à l'aide du schéma de Hörner pour la division (à partir de la fin):

$$\begin{aligned} y_1 &= 13 - x^2 : 15, \\ y_2 &= 11 - x^2 : y_1, \\ y_3 &= 9 - x^2 : y_2, \\ y_4 &= 7 - x^2 : y_3, \\ y_5 &= 5 - x^2 : y_4, \\ y_6 &= 3 - x^2 : y_5, \\ y = y_7 &= 1 - x^2 : y_6. \end{aligned}$$

Il en résulte que $\operatorname{tg} x = \frac{x}{y}$.

Exemple. Trouver tg 40°.

Solution. On a:

$$x = \text{arc } 40° = 0{,}698132$$

et

$$x^2 = 0{,}487388.$$

On en tire

$$y_1 = 13 - \frac{0{,}487388}{15} = 12{,}967508;$$

$$y_2 = 11 - \frac{0{,}487388}{12{,}967508} = 10{,}962413;$$

$$y_3 = 9 - \frac{0{,}487388}{10{,}962413} = 8{,}955540;$$

$$y_4 = 7 - \frac{0{,}487388}{8{,}955540} = 6{,}955577;$$

$$y_5 = 5 - \frac{0{,}487388}{6{,}955577} = 4{,}929928;$$

$$y_6 = 3 - \frac{0{,}487388}{4{,}929928} = 2{,}901137;$$

$$y = y_7 = 1 - \frac{0{,}487388}{2{,}901137} = 0{,}832001$$

et, par conséquent,

$$\operatorname{tg} 40^\circ = \frac{0,698132}{0,832001} = 0,839100.$$

Tous les chiffres du résultat obtenu sont exacts.

§ 9. Fonctions hyperboliques

A. Calcul des valeurs du sinus hyperbolique

On sait que

$$\operatorname{sh} x = \frac{e^x - e^{-x}}{2},$$

de plus

$$\operatorname{sh}(-x) = -\operatorname{sh} x.$$

Le sinus hyperbolique admet le développement

$$\operatorname{sh} x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots \quad (-\infty < x < +\infty).$$

En supposant que $x > 0$ il est commode de réaliser le calcul par le processus de sommation

$$\operatorname{sh} x = u_1 + u_2 + \dots + u_n + R_n,$$

où

$$u_1 = x, \quad u_{k+1} = \frac{x^2}{2k(2k+1)} u_k \quad (k = 1, 2, \dots, n-1)$$

et R_n est le reste. Avec $n \geqslant x > 0$ on a:

$$R_n = \frac{x^{2n+1}}{(2n+1)!} + \frac{x^{2n+3}}{(2n+3)!} + \frac{x^{2n+5}}{(2n+5)!} + \dots <$$

$$< \frac{x^{2n+1}}{(2n+1)!}\left[1 + \frac{x^2}{(2n+2)(2n+3)} + \frac{x^4}{(2n+2)^2(2n+3)^2} + \dots\right] <$$

$$< \frac{x^{2n+1}}{(2n+1)!} \cdot \frac{1}{1 - \frac{x^2}{(2n+2)(2n+3)}} < \frac{4}{3}\frac{x^{2n+1}}{(2n+1)!} = \frac{4}{3} u_{n+1}.$$

Comme il est évident que

$$u_{n+1} = \frac{x^2}{2n(2n+1)} u_n < \frac{1}{4} u_n,$$

on a

$$R_n < \frac{1}{3} u_n.$$

B. Calcul des valeurs du cosinus hyperbolique

On sait que

$$\operatorname{ch} x = \frac{e^x + e^{-x}}{2},$$

de plus,

$$\operatorname{ch}(-x) = \operatorname{ch} x.$$

Le cosinus hyperbolique donne lieu au développement

$$\operatorname{ch} x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots \quad (-\infty < x < +\infty).$$

Le calcul le plus commode se fait par sommation

$$\operatorname{ch} x = v_1 + v_2 + \ldots + v_n + R_n,$$

où

$$v_1 = 1, \quad v_{k+1} = \frac{x^2}{(2k-1)\,2k} v_k \quad (k = 1, 2, \ldots, n-1)$$

et R_n est le reste. Avec $n \geqslant |x| > 0$ on a :

$$R_n = \frac{x^{2n}}{(2n)!} + \frac{x^{2n+2}}{(2n+2)!} + \frac{x^{2n+4}}{(2n+4)!} + \ldots <$$

$$< \frac{x^{2n}}{(2n)!}\left[1 + \frac{x^2}{(2n+1)(2n+2)} + \frac{x^4}{(2n+1)^2(2n+2)^2} + \ldots\right] <$$

$$< \frac{x^{2n}}{(2n)!} \cdot \frac{1}{1 - \frac{x^2}{(2n+1)(2n+2)}} < \frac{4}{3} \cdot \frac{x^{2n}}{(2n)!} = \frac{4}{3} v_{n+1}.$$

Etant donné que l'inégalité

$$v_{n+1} = \frac{x^2}{(2n-1)\,2n} v_n \leqslant \frac{1}{2} v_n$$

se vérifie avec $n \geqslant 1$, on a

$$R_n < \frac{2}{3} v_n.$$

C. Calcul de la tangente hyperbolique

On a

$$\operatorname{th} x = \frac{\operatorname{sh} x}{\operatorname{ch} x} = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

où

$$\operatorname{th}(-x) = -\operatorname{th} x.$$

Pour de petits $|x|$ le calcul de la tangente hyperbolique se fait en utilisant le développement

$$\operatorname{th} x = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \frac{62x^9}{2835} + \ldots \left(|x| < \frac{\pi}{2}\right).$$

Quel que soit x, la valeur de la tangente hyperbolique s'exprime par la fraction continue

$$\operatorname{th} x = \left[0;\ \frac{x}{1},\ \frac{x^2}{3},\ \frac{x^2}{5},\ \ldots,\ \frac{x^2}{2n-1},\ \ldots\right],$$

et en outre, puisque avec $x > 0$ les éléments de cette fraction sont positifs, th x avec $x > 0$ est comprise entre deux fractions correspondantes voisines.

Si $x > 0$ est grand, il est commode de calculer th x en appliquant la formule

$$\operatorname{th} x = 1 - \frac{2}{e^{2x}+1}.$$

§ 10. Application de la méthode des itérations au calcul approché des fonctions

Supposons qu'il faille calculer la valeur de la fonction continue

$$y = f(x) \tag{1}$$

pour la valeur donnée de l'argument x. Si la fonction (1) est assez compliquée et impose le calcul d'un grand nombre de ses valeurs, les calculs se font généralement sur ordinateurs. Il se peut que les particularités fonctionelles de la machine rendent difficile le calcul immédiat des valeurs de la fonction à l'aide de la formule (1). Dans ces conditions les opérations les plus simples peuvent devenir compliquées et même irréalisables. Il existe, par exemple, des machines « sans division ». Des cas sont nombreux où il est alors avantageux d'appliquer l'artifice suivant. Ecrivons la fonction (1) sous une forme implicite

$$F(x, y) = 0. \tag{2}$$

Supposons que $F(x, y)$ est continue et a une dérivée partielle continue $F'_y(x, y) \neq 0$.

Soit y_n une valeur approchée de y. En appliquant le théorème de Lagrange, on a:

$$F(x, y_n) = F(x, y_n) - F(x, y) = (y_n - y)\, F'_y(x, \bar{y}_n),$$

où $\bar{y}_n$ est une valeur intermédiaire entre y_n et y. Il vient

$$y = y_n - \frac{F(x, y_n)}{F'_y(x, \bar{y}_n)}. \tag{3}$$

Nous ne connaissons pas la valeur $\bar{y}_n$. En posant $\bar{y}_n \approx y_n$, on obtient pour le calcul de la valeur de y le *processus itératif* [7]

$$y_{n+1} = y_n - \frac{F(x, y_n)}{F'_y(x, y_n)} \qquad (n = 0,\ 1,\ 2,\ \ldots). \tag{4}$$

L'interprétation géométrique de la formule (3) est bien simple. Fixons la valeur de x et considérons la courbe de la fonction

$$z = F(x, y). \qquad (4')$$

La formule (4) implique que le processus considéré est une application de la méthode de Newton (cf. chapitre IV, § 5) à la fonction (4), c'est-à-dire que les approximations successives de y_{n+1} s'obtiennent comme abscisses du point d'intersection avec l'axe Oy de la tangente à la courbe (4) tracée pour $y = y_n$ ($n = 0, 1, 2, \ldots$) (fig. 5). La convergence du processus sera assurée si les signes de

$$F'_y(x, y) \quad \text{et} \quad F''_{yy}(x, y)$$

sont constants dans l'intervalle considéré qui contient la racine y.

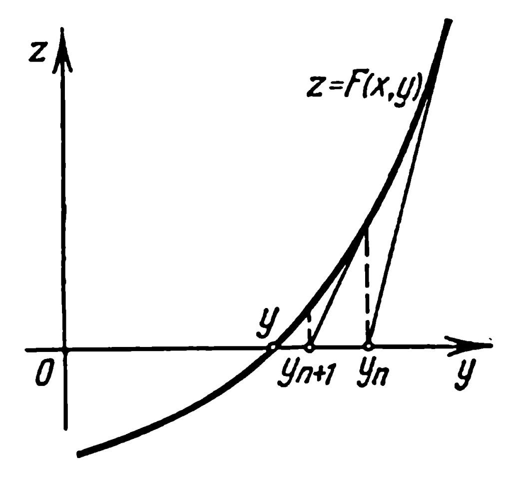


Fig. 5.

La valeur initiale y_0 est en général arbitraire et doit être choisie aussi proche que possible de la valeur recherchée de y. Le processus itératif se poursuit tant que dans les limites de la précision donnée ε deux valeurs successives y_n et y_{n-1} ne se confondent: $|y_{n-1} - y_n| < \varepsilon$. De plus, en toute rigueur, on ne garantit pas que

$$|y - y_n| < \varepsilon, \qquad (5)$$

et c'est pourquoi dans chaque cas concret il faut procéder à une exploration supplémentaire.

L'avantage que présentent les processus itératifs est l'uniformité des opérations et, par suite, la mise en programme relativement facile.

Notons que pour la fonction donnée (1) la représentation $F(x, y) = 0$ peut être réalisée d'une infinité de façons. Cette propriété doit être mise à profit pour obtenir un processus itératif convergeant rapidement. Dans les paragraphes qui suivent nous donnons les types des processus principaux.

§ 11. Calcul de la valeur inverse

Soit $y = \frac{1}{x}$.

Pour fixer les idées, considérons que $x > 0$. Posons

$$F(x, y) \equiv x - \frac{1}{y} = 0,$$

alors

$$F'_y(x, y) = \frac{1}{y^2}.$$

En appliquant la formule (4) du § 10 on a:

$$y_{n+1} = y_n - \frac{x - \frac{1}{y_n}}{\frac{1}{y_n^2}}$$

ou

$$y_{n+1} = y_n (2 - xy_n) \qquad (n = 0, 1, 2, \ldots), \tag{1}$$

c'est-à-dire nous obtenons un processus itératif sans division. La valeur initiale y_0 est choisie de la façon suivante. Soit l'argument x traduit en écriture binaire

$$x = 2^m x_1, \text{ où } m \text{ est un entier et } \frac{1}{2} \leqslant x_1 < 1.$$

On pose alors

$$y_0 = 2^{-m}. \tag{2}$$

Etablissons les conditions de convergence du processus (1). La formule (1) entraîne

$$\frac{1}{x} - y_n = \frac{1}{x} - 2y_{n-1} + xy_{n-1}^2 = x\left(\frac{1}{x} - y_{n-1}\right)^2; \tag{3}$$

d'où

$$\frac{1}{x} - y_n = x^{2^n - 1}\left(\frac{1}{x} - y_0\right)^{2^n} = \frac{1}{x}(1 - xy_0)^{2^n}. \tag{4}$$

Pour assurer la convergence du processus (4) il faut et il suffit de vérifier l'inégalité

$$|1 - xy_0| < 1$$

ou

$$-1 < 1 - xy_0 < 1.$$

Finalement on obtient le résultat suivant: si

$$0 < xy_0 < 2, \tag{5}$$

on a

$$\lim_{n \to \infty} y_n = \frac{1}{x}$$

Notons que notre choix de y_0 (2) entraîne

$$xy_0 = 2^m x_1 \cdot 2^{-m} = x_1;$$

donc

$$\frac{1}{2} \leqslant xy_0 < 1, \tag{6}$$

par conséquent, la condition (5) est justifiée. En outre, la formule (3) conduit à

$$\left|\frac{1}{x}-y_n\right| \leqslant \frac{1}{x}\left(\frac{1}{2}\right)^{2^n} \leqslant 2y_0\left(\frac{1}{2}\right)^{2^n},$$

c'est-à-dire la convergence du processus itératif est extrêmement rapide.

Parfois il est plus commode en pratique d'utiliser une autre estimation de l'erreur de la valeur de y_n. Constatons d'abord que dans le cas considéré les approximations successives $y_0, y_1, y_2, \ldots$ s'obtiennent par la méthode de Newton appliquée à l'hyperbole

$$z = x - \frac{1}{y} \qquad (x = \text{const})$$

(fig. 6). L'inégalité (6) et la formule (3) donnent

$$0 < y_n < \frac{1}{x} \quad (n = 0, 1, 2, \ldots).$$

D'autre part, puisque

$$y_n - y_{n-1} = y_{n-1}(1 - xy_{n-1}) = \\ = xy_{n-1}\left(\frac{1}{x} - y_{n-1}\right) \geqslant 0, \qquad (7)$$

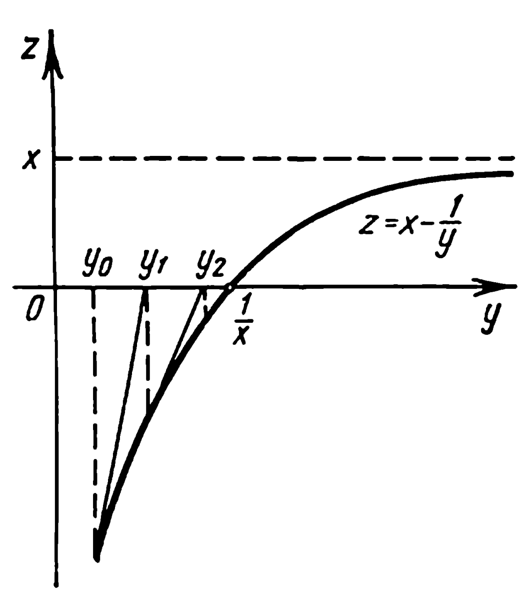


Fig. 6.

les approximations successives de y_n sont croissantes monotones:

$$y_0 \leqslant y_1 \leqslant y_2 \leqslant \ldots$$

La formule (7) amène

$$\frac{1}{x} - y_{n-1} = \frac{1}{xy_{n-1}}(y_n - y_{n-1}),$$

ou, du fait que

$$xy_{n-1} \geqslant xy_0 \geqslant \frac{1}{2},$$

on a

$$\frac{1}{x} - y_{n-1} \leqslant 2(y_n - y_{n-1}).$$

On en tire

$$\frac{1}{x} - y_n \leqslant y_n - y_{n-1}.$$

Par conséquent, si l'on établit que $y_n - y_{n-1} < \varepsilon$, l'erreur vraie est également

$$0 < \frac{1}{x} - y_n < \varepsilon.$$

E x e m p l e. Chercher à l'aide de (1) la valeur de la fonction $y=\frac{1}{x}$ avec $x=3$.

S o l u t i o n. Ici $x=2^2\cdot\frac{3}{4}$. Posons $y_0=\frac{1}{4}$, on a alors

$$y_1=\frac{1}{4}\left(2-\frac{3}{4}\right)=\frac{5}{16}=0{,}312;$$

$$y_2=0{,}312\,(2-3\cdot 0{,}312)=0{,}332, \text{ etc.}$$

Le processus itératif converge rapidement.

§ 12. Racine carrée

Soit

$$y=\sqrt{x} \qquad (x>0). \tag{1}$$

Posons

$$F(x,\ y)\equiv y^2-x=0,$$

alors

$$F'_y(x,\ y)=2y.$$

En appliquant la formule (4) du § 10, on a le *processus de Héron* :

$$y_{n+1}=y_n-\frac{y_n^2-x}{2y_n}$$

ou

$$y_{n+1}=\frac{1}{2}\left(y_n+\frac{x}{y_n}\right) \tag{2}$$

$$(n=0,\ 1,\ 2,\ \ldots).$$

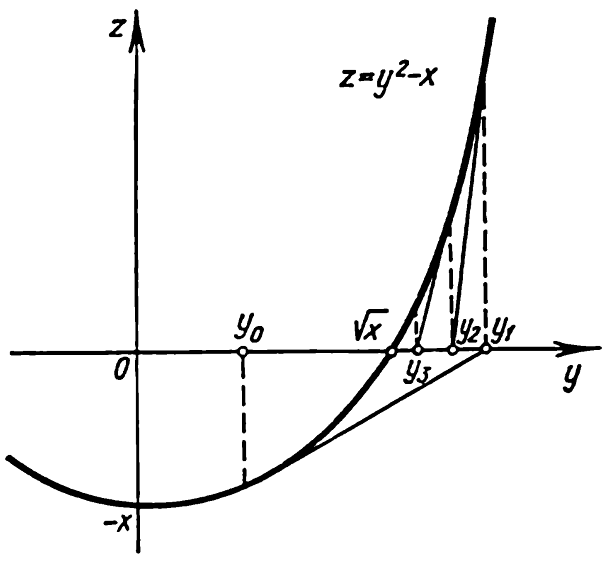


Fig. 7.

Les approximations successives y_0, y_1, y_2, . . . s'obtiennent évidemment d'après la méthode de Newton appliquée à la parabole

$$z=y^2-x \qquad (x=\text{const})$$

(fig. 7).

Remarquons que si l'on prend pour y_0 la valeur tabulée qui donne $\sqrt{x}$ avec une erreur relative $|\,\delta\,|$, alors y_1 définie d'après la formule (2) donnera la valeur de $\sqrt{x}$ avec une erreur relative approchée $\frac{1}{2}\,\delta^2$.

En effet, en posant

$$y_0=\sqrt{x}\,(1+\delta)$$

et en négligeant les puissances de δ supérieures à 3, on a :

$$y_1 = \frac{1}{2}\left(y_0 + \frac{x}{y_0}\right) = \frac{1}{2}[\sqrt{x}(1+\delta) + \sqrt{x}(1+\delta)^{-1}] =$$
$$= \frac{1}{2}\sqrt{x}(1+\delta+1-\delta+\delta^2) = \sqrt{x}\left(1+\frac{\delta^2}{2}\right).$$

On en déduit la conclusion importante: *en appliquant le processus de Héron, le nombre de chiffres exacts devient à chaque pas à peu près deux fois plus grand que leur nombre antérieur.*

Exemple 1. Pour $y = \sqrt{2}$ on a approximativement :

$$y_0 = 1{,}4.$$

En améliorant la précision de cette valeur on a

$$y_1 = \frac{1}{2}\left(1{,}4 + \frac{2}{1{,}4}\right) = 0{,}7 + 0{,}714 = 1{,}414.$$

En reprenant le processus, on obtient:

$$y_2 = \frac{1}{2}\left(1{,}414 + \frac{2}{1{,}414}\right) = 0{,}707 + 0{,}7072136 = 1{,}4142136,$$

huit ou sept décimales étant exactes. En effet

$$\sqrt{2} = 1{,}41421356\ldots$$

Etablissons les conditions de convergence du processus de Héron. Si l'on remplace $n+1$ par n avec $y_0 \neq 0$, la formule (2) conduit à

$$y_n - \sqrt{x} = \frac{1}{2y_{n-1}}(y_{n-1} - \sqrt{x})^2$$

et

$$y_n + \sqrt{x} = \frac{1}{2y_{n-1}}(y_{n-1} + \sqrt{x})^2.$$

Il en résulte que

$$\frac{y_n - \sqrt{x}}{y_n + \sqrt{x}} = \left(\frac{y_{n-1} - \sqrt{x}}{y_{n-1} + \sqrt{x}}\right)^2. \tag{3}$$

Par conséquent,

$$\frac{y_n - \sqrt{x}}{y_n + \sqrt{x}} = \left(\frac{y_0 - \sqrt{x}}{y_0 + \sqrt{x}}\right)^{2^n}$$

et

$$y_n - \sqrt{x} = 2\sqrt{x} \cdot \frac{q^{2^n}}{1 - q^{2^n}}, \tag{4}$$

où

$$q = \frac{y_0 - \sqrt{x}}{y_0 + \sqrt{x}}. \tag{5}$$

La formule (4) entraîne que le processus de Héron converge avec

$$|q| < 1,$$

c'est-à-dire si

$$y_0 > 0.$$

Dans ce cas on a évidemment:

$$\lim_{n\to\infty} y_n = \sqrt{x}$$

et

$$y_n \geqslant \sqrt{x} \qquad (n = 1, 2, \ldots).$$

Constatons que

$$y_{n-1} - y_n = y_{n-1} - \frac{1}{2}\left(y_{n-1} + \frac{x}{y_{n-1}}\right) = \frac{y_{n-1}^2 - x}{2y_{n-1}} > 0, \qquad (6)$$

c'est pourquoi les approximations y_n avec $n \geqslant 1$ forment une suite décroissante

$$y_1 \geqslant y_2 \geqslant \ldots \geqslant y_{n-1} \geqslant y_n \geqslant \ldots \geqslant \sqrt{x}\,^{*}.$$

Pour travailler sur un calculateur électronique il est commode de mettre le nombre x sous une forme binaire

$$x = 2^m x_1, \quad \text{où } m \text{ est entier et } \frac{1}{2} \leqslant x_1 < 1.$$

On adopte alors, généralement, comme approximation initiale

$$y_0 = 2^{E\left(\frac{m}{2}\right)}, \qquad (7)$$

où $E\left(\frac{m}{2}\right)$ est la partie entière du nombre $\frac{m}{2}$.

Exemple 2. Trouver $\sqrt{5}$.

Solution. Ici $x = 5 = 2^3 \cdot \frac{5}{8}$. Donc

$$y_0 = 2^{E\left(\frac{3}{2}\right)} = 2.$$

D'après la formule (2) on a successivement:

$$y_1 = \frac{1}{2}\left(2 + \frac{5}{2}\right) = 2{,}25,$$

$$y_2 = \frac{1}{2}\left(2{,}25 + \frac{5}{2{,}25}\right) = \frac{1}{2}(2{,}25 + 2{,}2222) = 2{,}2361,$$

etc. La table des racines carrées donne:

$$\sqrt{5} = 2{,}236068\ldots$$

---

* Le signe d'égalité ne peut avoir lieu que si $y_0 = \sqrt{x}$.

Evaluons la quantité $|q|$ exprimée par la formule (5) d'après la valeur y_0 définie par la formule (7).

Si $m = 2p$ est un nombre pair, on a:

$$y_0 = 2^{E\left(\frac{m}{2}\right)} = 2^p > \sqrt{x}$$

et, par conséquent,

$$|q| = \frac{y_0 - \sqrt{x}}{y_0 + \sqrt{x}} = \frac{2^p - 2^p\sqrt{x_1}}{2^p + 2^p\sqrt{x_1}} = \frac{1 - \sqrt{x_1}}{1 + \sqrt{x_1}} \leqslant \frac{1 - \sqrt{\frac{1}{2}}}{1 + \sqrt{\frac{1}{2}}} = (\sqrt{2} - 1)^2.$$

D'une façon analogue, si $m = 2p + 1$ est impair,

$$y_0 = 2^{E\left(\frac{m}{2}\right)} = 2^p \leqslant \sqrt{x}.$$

C'est pourquoi

$$|q| = \frac{\sqrt{x} - y_0}{\sqrt{x} + y_0} = \frac{2^p\sqrt{2x_1} - 2^p}{2^p\sqrt{2x_1} + 2^p} = \frac{\sqrt{2x_1} - 1}{\sqrt{2x_1} + 1} =$$

$$= 1 - \frac{2}{\sqrt{2x_1} + 1} < 1 - \frac{2}{\sqrt{2} + 1} = (\sqrt{2} - 1)^2.$$

Ainsi on a toujours:

$$|q| \leqslant (\sqrt{2} - 1)^2 = 0{,}1716\ldots < \frac{1}{5}.$$

Il en résulte en vertu de la formule (4):

$$0 \leqslant y_n - \sqrt{x} < 2\sqrt{x} \cdot \frac{\left(\frac{1}{5}\right)^{2^n}}{1 - \left(\frac{1}{5}\right)^{2^n}} \leqslant \frac{25}{12} y_1 \left(\frac{1}{5}\right)^{2^n} \text{ avec } n \geqslant 1,$$

où

$$y_1 = \frac{1}{2}\left(y_0 + \frac{x}{y_0}\right) \leqslant \frac{3}{2} y_0.$$

Il s'ensuit que

$$0 \leqslant y_n - \sqrt{x} < \frac{25}{8} y_0 \left(\frac{1}{5}\right)^{2^n}. \tag{8}$$

La formule (8) permet de définir aisément le nombre d'itérations $n = n(x)$ suffisant pour assurer la précision imposée.

Voici encore une formule pour évaluer l'erreur de la valeur y_n ($n \geqslant 2$). Etant donné que

$$y_{n-1} \geqslant \sqrt{x} \text{ et } \frac{x}{y_{n-1}} \leqslant \sqrt{x}$$

et tenant compte de la formule (6), on a:

$$y_{n-1} - \sqrt{x} \leqslant y_{n-1} - \frac{x}{y_{n-1}} = \frac{y_{n-1}^2 - x}{y_{n-1}} = 2(y_{n-1} - y_n).$$

Par conséquent,

$$0 \leqslant y_n - \sqrt{x} \leqslant y_{n-1} - y_n. \qquad (9)$$

Donc, si $0 \leqslant y_{n-1} - y_n < \varepsilon \, (n \geqslant 2)$, on assure que $0 \leqslant y_n - \sqrt{x} < \varepsilon$.

Voici encore un procédé pour calculer la racine carrée, quelquefois très utile. Remplaçons la fonction (1) par une relation équivalente

$$F(x, y) \equiv \frac{x}{y^2} - 1 = 0.$$

Il vient

$$F'_y(x, y) = -\frac{2x}{y^3}.$$

En appliquant la formule (4) du § 10, on a:

$$y_{n+1} = y_n + \frac{\frac{x}{y_n^2} - 1}{\frac{2x}{y_n^3}}$$

ou

$$y_{n+1} = \frac{y_n}{2}\left(3 - \frac{y_n^2}{x}\right) \quad (n = 0, 1, 2, \ldots). \qquad (10)$$

Nous omettons l'étude des conditions de convergence du processus itératif (10) et l'estimation de l'érreur.

§ 13. Valeur inverse de la racine carrée

Posons

$$y = \frac{1}{\sqrt{x}} \qquad (x > 0).$$

Si l'on met cette fonction sous la forme

$$y = \sqrt{\frac{1}{x}},$$

la formule (10) du paragraphe précédent permet d'obtenir le processus itératif «sans division»

$$y_{n+1} = \frac{y_n}{2}(3 - x y_n^2) \qquad (n = 0, 1, 2, \ldots). \qquad (1)$$

Si $x = 2^m x_1$, où $\frac{1}{2} \leqslant x_1 < 1$, on prend pour y_0 la valeur

$$y_0 = 2^{-E\left(\frac{m}{2}\right)}.$$

Notons qu'en faisant appel à l'égalité évidente

$$\sqrt{x} = x\sqrt{\frac{1}{x}},$$

on rend également possible, en vertu de la formule (1), l'extraction de la racine carrée d'un nombre « sans division ».

§ 14. Racine cubique

Si

$$y = \sqrt[3]{x} \quad (x > 0), \qquad (1)$$

en posant

$$F(x, y) \equiv y^3 - x = 0,$$

on a:

$$F'_y(x, y) = 3y^2.$$

L'utilisation de la formule (4) du § 10 conduit à:

$$y_{n+1} = y_n - \frac{y_n^3 - x}{3y_n^2} \qquad (2)$$

ou

$$y_{n+1} = \frac{1}{3}\left(2y_n + \frac{x}{y_n^2}\right). \qquad (3)$$

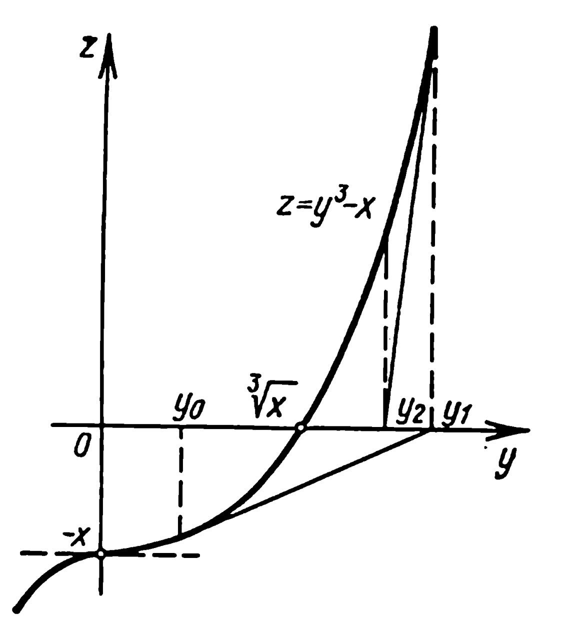


Fig. 8.

L'interprétation géométrique du processus (3) est donnée par la méthode de Newton appliquée à la parabole cubique

$$z = y^3 - x \quad (x = \text{const})$$

(fig. 8). Le processus (3) converge avec $y_0 > 0$.

Si l'on prend comme approximation initiale y_0 la valeur tabulée de $\sqrt[3]{x}$ avec une erreur relative $|\delta|$, c'est-à-dire si l'on pose

$$y_0 = \sqrt[3]{x}(1 + \delta),$$

la valeur y_1, fournie par la formule (3), donne $\sqrt[3]{x}$ avec une erreur relative δ^2. En effet, en utilisant la formule (3), on a:

$$y_1 = \frac{1}{3}\left(2y_0 + \frac{x}{y_0^2}\right) = \frac{1}{3}\left[2\sqrt[3]{x}(1+\delta) + \sqrt[3]{x}(1+\delta)^{-2}\right] =$$

$$= \frac{1}{3}\sqrt[3]{x}(2 + 2\delta + 1 - 2\delta + 3\delta^2) = \sqrt[3]{x}(1 + \delta^2).$$

On en tire, en particulier, que si y_0 compte p chiffres exacts au sens strict, y_1 aura à peu près $2p$ ou $2p - 1$ chiffres exacts au sens lâche (cf. § 12).

E x e m p l e. Les tables à trois décimales donnent:

$$\sqrt[3]{10}=2{,}154,$$

où tous les chiffres sont exacts.

En utilisant la formule (3) on obtient

$$\sqrt[3]{10}=\frac{1}{3}\left(2\cdot 2{,}154+\frac{10}{2{,}154^2}\right)=\frac{1}{3}(2\cdot 2{,}154+2{,}155304)=2{,}154435.$$

A titre de comparaison, voici la valeur tirée de la table de Burlow

$$\sqrt[3]{10}=2{,}1544347\ \ldots$$

Si $x = 2^m x_1$, où m est un entier et $\frac{1}{2}\leqslant x_1<1$, on choisit en général comme valeur initiale y_0

$$y_0=2^{E\left(\frac{m}{3}\right)}>0. \tag{4}$$

Puisque

$$y_n-\sqrt[3]{x}=\frac{1}{3}\left(2y_{n-1}+\frac{x}{y_{n-1}^2}-3\sqrt[3]{x}\right)=$$
$$=\frac{1}{3y_{n-1}^2}\left(y_{n-1}-\sqrt[3]{x}\right)^2\left(2y_{n-1}+\sqrt[3]{x}\right)>0,$$

il vient

$$y_n\geqslant\sqrt[3]{x} \text{ avec } n\geqslant 1. \tag{5}$$

De plus, en remplaçant $n + 1$ par n dans la formule (2), on a

$$y_{n-1}-y_n=\frac{y_{n-1}^3-x}{3y_{n-1}^2}; \tag{6}$$

donc

$$y_1\geqslant y_2\geqslant\ldots\geqslant y_{n-1}\geqslant y_n\geqslant\ldots\geqslant\sqrt[3]{x}. \tag{7}$$

Il existe donc une limite

$$\lim_{n\to\infty} y_n=y>0.$$

En passant dans l'égalité (3) à la limite quand $n\to\infty$, on obtient:

$$y=\frac{1}{3}\left(2y+\frac{x}{y^2}\right),$$

c'est-à-dire $y^3 = x$ et, par conséquent, $y=\sqrt[3]{x}$. Ainsi

$$\lim_{n\to\infty} y_n=\sqrt[3]{x}.$$

Si l'approximation initiale y_0 est choisie d'après la formule (4), on peut montrer qu'avec $n\geqslant 2$

$$0\leqslant y_n-\sqrt[3]{x}\leqslant\frac{3}{2}(y_{n-1}-y_n).$$

BIBLIOGRAPHIE

1. *V. Smirnov*. Cours de mathématiques supérieures, t. I. Editions Mir, Moscou, 1969, chapitre IV.
2. *A. Markov*. Calcul des différences finies. 2e éd., Matézis, 1911, chapitre III.
3. *G. Tolstov*. Cours d'analyse mathématique, t. II. Gostekhizdat, Moscou, 1957, chapitre XXIV.
4. *A. Khovanski*. Applications des fractions continues et de leurs généralisations aux problèmes d'analyse approchée. Gostekhizdat, 1956, chapitre II.
5. *B. Kagan* et *T. Ter-Mikaélian*. Résolution des problèmes d'ingénieur sur les calculateurs digitaux. Gosénergoïzdat, Moscou-Léningrad, 1958, chapitre III.
6. *G. Fichtengoltz*. Cours de calcul différentiel et intégral. OGIZ, Moscou-Léningrad, 1948, t. II, chapitre XII.
7. *L. Lusternik*, *A. Abramov*, *V. Chestakov*, *M. Choura-Boura*. Résolution des problèmes mathématiques sur les calculateurs digitaux. Editions de l'Académie des Sciences de l'U.R.S.S., 1952.

CHAPITRE IV

RÉSOLUTION APPROCHÉE DES ÉQUATIONS ALGÉBRIQUES ET TRANSCENDANTES

§ 1. Séparation des racines

Si une équation algébrique ou transcendante est suffisamment complexe, il est relativement rare qu'on puisse obtenir ses racines avec précision. Par ailleurs, dans certains cas les coefficients de l'équation ne sont connus qu'approximativement et, par conséquent, le problème de la détermination précise des racines proprement dit perd son sens. C'est pourquoi les méthodes de la recherche approchée des racines d'une équation et l'estimation du degré de sa précision acquièrent un intérêt particulier.

Soit l'équation

$$f(x) = 0, \tag{1}$$

où la fonction $f(x)$ est définie et continue dans un certain intervalle fini ou infini $a < x < b$.

Dans ce qui suit nous devrons recourir à la dérivée première $f'(x)$ et même à la dérivée seconde $f''(x)$, ce que nous stipulerons alors spécialement.

Toute valeur ξ qui annule la fonction $f(x)$, c'est-à-dire telle que

$$f(\xi) = 0,$$

s'appelle *racine de l'équation* (1) ou *zéro* de la fonction $f(x)$.

Nous supposons que l'équation (1) ne possède que des *racines isolées*, c'est-à-dire que pour chaque racine de l'équation (1) il existe un voisinage qui ne contient pas d'autres racines de cette équation.

La recherche approchée des racines réelles isolées de l'équation (1) se fait en général en deux étapes:

1) séparation des racines, qui consiste à établir des segments $[\alpha, \beta]$ les plus serrés possibles contenant une et seulement une racine de l'équation (1);

2) amélioration de la précision ou mise au point des racines approchées, c'est-à-dire l'obtention de leur précision imposée.

Pour réaliser la séparation des racines on fait appel au théorème connu de l'analyse mathématique ([5], chapitre IV).

Théorème 1. *Si une fonction continue $f(x)$ prend aux extrémités du segment $[\alpha, \beta]$ des valeurs de signes contraires, c'est-à-dire si $f(\alpha) f(\beta) < 0$, ce segment contient au moins une racine de l'équation $f(x) = 0$, fait qui traduit l'existence au moins d'un nombre $\xi \in [\alpha, \beta]$* * *tel que $f(\xi) = 0$* (fig. 9).

Si la dérivée $f'(x)$ existe et garde son signe constant dans l'intervalle (α, β), c'est-à-dire si $f'(x) > 0$ (ou $f'(x) < 0$) avec $\alpha < x < \beta$ (fig. 10), la racine ξ est unique.

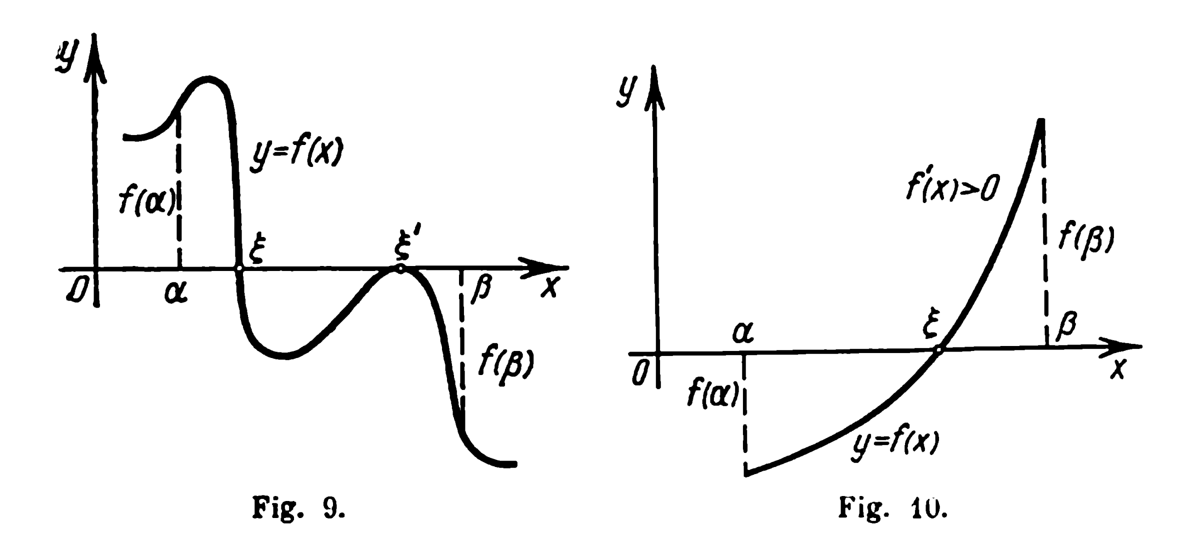


Fig. 9. Fig. 10.

Le processus de séparation des racines débute par la détermination des signes de la fonction $f(x)$ aux points frontières $x = a$ et $x = b$ du domaine de son existence.

Puis on détermine les signes de la fonction $f(x)$ en quelques points intermédiaires $x = \alpha_1, \alpha_2, \ldots$, dont le choix rend compte des particularités de la fonction $f(x)$. S'il se trouve que $f(\alpha_k) f(\alpha_{k+1}) < 0$, en vertu du théorème 1, dans l'intervalle (α_k, α_{k+1}) l'équation $f(x) = 0$ possède une racine. Il faut alors définir d une façon ou d'une autre si cette racine est unique. Pour séparer les racines, il suffit souvent en pratique d'opérer une bipartition qui consiste à diviser approximativement l'intervalle donné (α, β) en deux, quatre, huit, etc., parties égales (jusqu'à un certain pas) et à déterminer les signes de la fonction $f(x)$ aux points de division. Il est utile de retenir que l'équation algébrique de degré n

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0 \quad (a_0 \neq 0)$$

compte au plus n racines réelles. Donc, si nous avons obtenu pour une telle équation $n + 1$ changements de signes, toutes ses racines sont alors séparées.

---

* La notation $\xi \in (\alpha, \beta)$ signifie que le point ξ appartient à l'intervalle (α, β).

Exemple 1. Séparer les racines de l'équation

$$f(x) \equiv x^3 - 6x + 2 = 0. \qquad (2)$$

Solution. Composons un schéma approché

x	$f(x)$	x	$f(x)$
$-\infty$	$-$	1	$-$
-3	$-$	3	$+$
-1	$+$	$+\infty$	$+$
0	$+$		

Par conséquent, l'équation (2) possède trois racines réelles comprises dans les intervalles $(-3, -1)$, $(0, 1)$ et $(1, 3)$.

S'il existe une dérivée continue $f'(x)$ et les racines de l'équation

$$f'(x) = 0$$

se calculent aisément, le processus de séparation des racines peut être ordonné. Il est clair qu'à cet effet il suffit de compter les signes de la fonction $f(x)$ aux points des zéros de sa dérivée et aux points frontières $x = a$ et $x = b$.

Exemple 2. Séparer les racines de l'équation

$$f(x) \equiv x^4 - 4x - 1 = 0. \qquad (3)$$

Solution. Ici $f'(x) = 4(x^3 - 1)$, d'où $f'(x) = 0$ avec $x = 1$.

On a $f(-\infty) > 0\ (+)$; $f(1) < 0\ (-)$; $f(+\infty) > 0\ (+)$. Par conséquent, l'équation (3) n'admet que deux racines réelles dont une repose dans l'intervalle $(-\infty, 1)$ et l'autre dans l'intervalle $(1, +\infty)$.

Exemple 3. Déterminer le nombre de racines réelles de l'équation

$$f(x) \equiv x + e^x = 0. \qquad (4)$$

Solution. Etant donné que $f'(x) = 1 + e^x > 0$ et $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, l'équation (4) n'a qu'une seule racine réelle.

Evaluons maintenant l'erreur d'une racine approchée.

Théorème 2. *Soit ξ une racine exacte et $\bar{x}$ une racine approchée de l'équation $f(x) = 0$, qui reposent sur le même segment $[\alpha, \beta]$, de plus $|f'(x)| \geqslant m_1 > 0$ pour $\alpha \leqslant x \leqslant \beta$* *.

---

* On peut prendre notamment pour m_1 la valeur minimale de $|f'(x)|$ avec $\alpha \leqslant x \leqslant \beta$.

Dans ces conditions l'estimation valable s'écrit

$$|\bar{x}-\xi| \leqslant \frac{|f(\bar{x})|}{m_1}. \tag{5}$$

D é m o n s t r a t i o n. En appliquant le théorème de Lagrange, on a :

$$f(\bar{x})-f(\xi)=(\bar{x}-\xi)f'(c),$$

où c est une valeur intermédiaire entre $\bar{x}$ et ξ, ou $c \in (\alpha, \beta)$.

Etant donné que $f(\xi) = 0$ et $|f'(c)| \geqslant m_1$, il en résulte que

$$|f(\bar{x})-f(\xi)|=|f(\bar{x})| \geqslant m_1 |\bar{x}-\xi|.$$

Par conséquent,

$$|\bar{x}-\xi| \geqslant \frac{|f(\bar{x})|}{m_1}.$$

R e m a r q u e. Les résultats fournis par la formule (5) peuvent être grossiers et son application n'est pas toujours commode. C'est pourquoi en pratique on réduit par tel ou tel procédé l'intervalle commun (α, β), qui contient la racine ξ et sa valeur approchée $\bar{x}$, et on pose $|\bar{x} - \xi| \leqslant \beta - \alpha$.

E x e m p l e 4. Une racine approchée de l'équation $f(x) \equiv \equiv x^4 - x - 1 = 0$ est $\bar{x} = 1{,}22$. Evaluer l'erreur absolue de cette racine.

S o l u t i o n. On a $f(\bar{x}) = 2{,}2153 - 1{,}22 - 1 = -0{,}0047$.

Etant donné qu'avec $\bar{\bar{x}} = 1{,}23$ on obtient

$$f(\bar{\bar{x}}) = 2{,}2888 - 1{,}23 - 1 = +0{,}0588,$$

la racine exacte ξ est comprise dans l'intervalle $(1{,}22;\ 1{,}23)$. La dérivée $f'(x) = 3x^3 - 1$ est croissante. C'est pourquoi dans l'intervalle considéré sa valeur minimale s'écrit :

$$m_1 = 3 \cdot 1{,}22^3 - 1 = 3 \cdot 1{,}816 - 1 = 4{,}448.$$

Il en résulte d'après la formule (5) :

$$|\bar{x}-\xi| \geqslant \frac{0{,}0047}{4{,}448} \approx 0{,}001.$$

R e m a r q u e. En pratique la précision d'une racine approchée $\bar{x}$ est parfois évaluée suivant qu'elle vérifie bien ou mal l'équation donnée $f(x) = 0$, c'est-à-dire si le nombre $|f(\bar{x})|$ est petit, on considère que $\bar{x}$ est une bonne approximation de la racine exacte ξ ; mais si $|f(\bar{x})|$ est grand, on admet que $\bar{x}$ est une valeur grossière de la

racine exacte ξ. Comme le montrent les figures 11 et 12, une telle attitude est incorrecte. Il ne faut pas non plus oublier que si l'on multiplie l'équation $f(x) = 0$ par un nombre arbitraire $N \neq 0$,

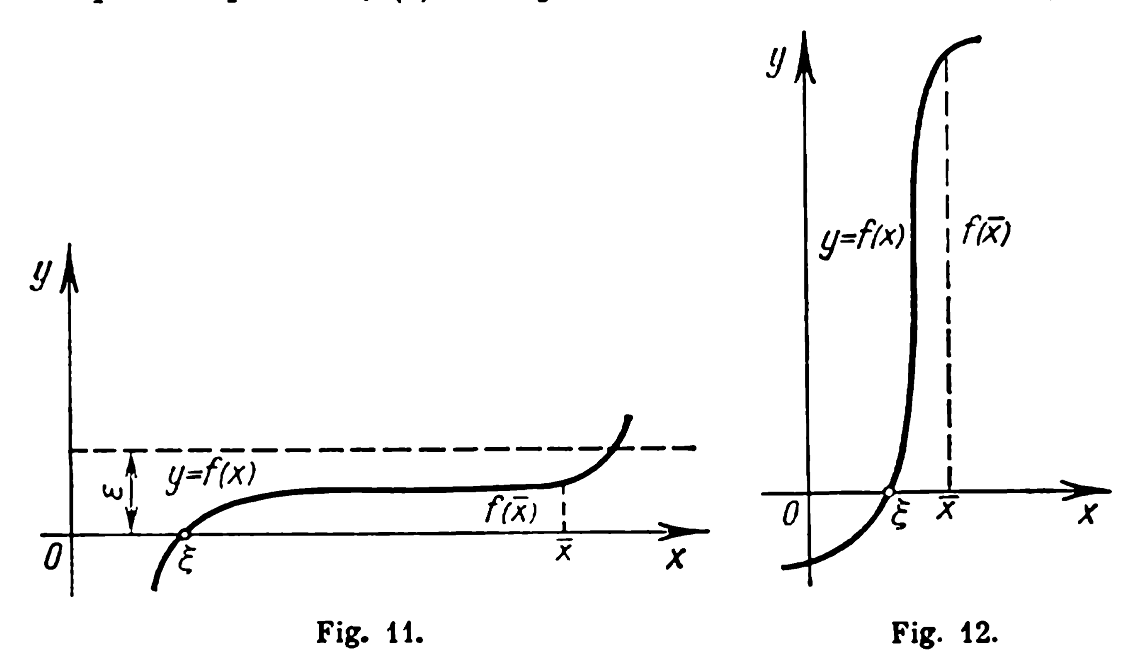


Fig. 11. Fig. 12.

on obtient une équation équivalente $Nf(x) = 0$, le nombre $|Nf(\bar{x})|$ pouvant être aussi grand ou aussi petit que l'on veut selon le choix du facteur N.

§ 2. Résolution graphique des équations

Les racines réelles d'une équation

$$f(x) = 0 \tag{1}$$

peuvent être déterminées approximativement comme les abscisses des points d'intersection de la courbe de la fonction $y = f(x)$ avec l'axe Ox (fig. 9). Si l'équation (1) n'admet pas de racines voisines, ce procédé permet de les séparer sans peine. Il arrive souvent qu il est avantageux de remplacer l'équation (1) par une équation équivalente *

$$\varphi(x) = \psi(x), \tag{2}$$

où les fonctions $\varphi(x)$ et $\psi(x)$ sont plus simples que $f(x)$. En construisant alors les courbes des fonctions $y = \varphi(x)$ et $y = \psi(x)$ on obtient les racines cherchées comme les abscisses des points d'intersection de ces courbes.

* Deux équations sont équivalentes si toutes leurs racines sont les mêmes.

Exemple 1. Résoudre graphiquement l'équation

$$x \lg x = 1. \qquad (3)$$

Solution. Mettons l'équation (3) sous la forme de l'égalité

$$\lg x = \frac{1}{x}.$$

On voit bien que les racines de l'équation (3) peuvent être définies comme les abscisses des points d'intersection de la courbe logarithmique $y = \lg x$ et de l'hyperbole $y = \frac{1}{x}$. La construction de ces courbes (fig. 13) sur du papier quadrillé fournit approximativement la racine unique $\xi \approx 2,5$ de l'équation (3).

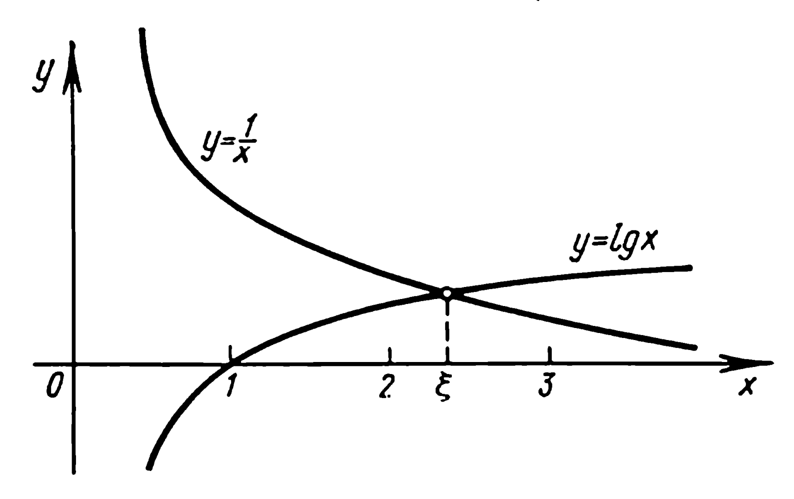


Fig. 13.

La recherche des racines de l'équation (3) devient plus simple si l'une des fonctions $\varphi(x)$ ou $\psi(x)$ est linéaire, c'est-à-dire si, par exemple, $\varphi(x) = ax + b$. Dans ce cas les racines de l'équation (2) s'obtiennent comme les abscisses des points d'intersection de la courbe $y = \psi(x)$ et de la droite $y = ax + b$. L'intérêt de ce procédé est surtout grand quand il faut résoudre plusieurs équations de même type qui ne diffèrent que par les coefficients a et b de la fonction linéaire. La construction graphique se ramène ici à la recherche des points d'intersection de la courbe fixée $y = \psi(x)$ avec des droites différentes. Les équations

$$x^n + ax + b = 0$$

se rapportent évidemment au type considéré.

Exemple 2. Résoudre les équations cubiques

$$x^3 - 1,75x + 0,75 = 0$$

et

$$x^3 + 2x + 7,8 = 0.$$

S o l u t i o n. Construisons une parabole cubique $y = x^3$. Les racines cherchées s'obtiennent comme les points d'intersection de cette parabole avec les droites (fig. 14) $y = 1,75x - 0,75$ et $y = -2x - 7,8$. Le dessin montre clairement que la première équation possède trois racines réelles: $x_1 = -1,5$; $x_2 = 0,5$; $x_3 = 1$, et la deuxième n'en compte qu'une: $x_1 = -1,65$.

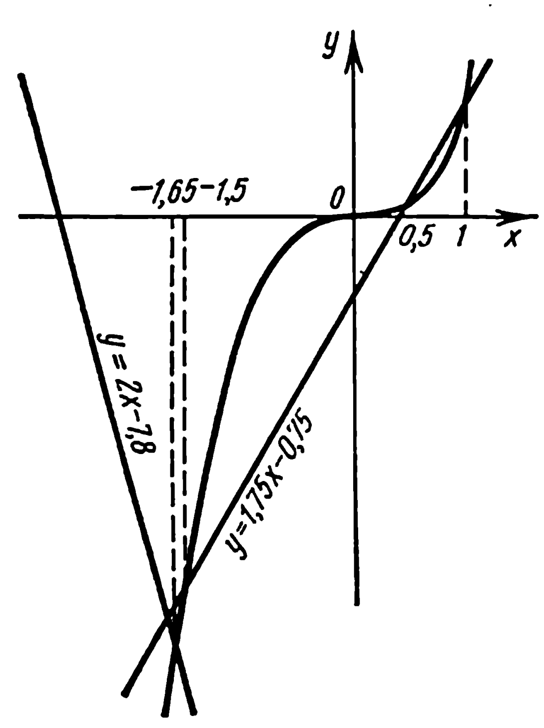


Fig. 14.

Bien que les méthodes graphiques de résolution des équations soient très commodes et relativement simples, elles ne sont généralement applicables qu'à une estimation grossière des racines. Le cas particulièrement défavorable au sens de manque de précision est celui des lignes qui se coupent sous un angle très aigu et qui pratiquement se confondent suivant un certain arc.

Les *méthodes nomographiques* ou des *abaques* étant une variante des méthodes graphiques, nous adressons le lecteur désireux de s'y initier aux ouvrages appropriés.

§ 3. Méthode de bipartition

Soit l'équation

$$f(x) = 0, \tag{1}$$

où la fonction $f(x)$ est continue sur $[a, b]$ et $f(a) f(b) < 0$.

Pour chercher la racine de l'équation (1) qui appartient au segment $[a, b]$, divisons ce segment en deux. Si $f\left(\frac{a+b}{2}\right) = 0$, $\xi = \frac{a+b}{2}$ est une racine de l'équation. Si $f\left(\frac{a+b}{2}\right) \neq 0$, prenons celle des moitiés $\left[a, \frac{a+b}{2}\right]$ ou $\left[\frac{a+b}{2}, b\right]$ aux extrémités de laquelle la fonction $f(x)$ a des signes opposés. Le nouveau segment raccourci $[a_1, b_1]$ est encore partitionné en deux, après quoi on reprend le raisonnement ci-dessus. On obtient ainsi à une certaine étape soit une racine exacte de l'équation (1), soit une suite infinie de segments emboîtés $[a_1, b_1]$, $[a_2, b_2]$, . . ., $[a_n, b_n]$, . . . tels que

$$f(a_n) f(b_n) < 0 \qquad (n = 1, 2, \ldots) \tag{2}$$

et

$$b_n - a_n = \frac{1}{2^n}(b-a). \tag{3}$$

Les extrémités gauches $a_1, a_2, \ldots, a_n, \ldots$ formant une suite non décroissante bornée et les extrémités droites $b_1, b_2, \ldots, b_n, \ldots$ une suite non croissante bornée, l'égalité (3) donne lieu à une limite commune

$$\xi = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

En passant dans l'inégalité (2) à la limite pour $n \to \infty$, la continuité de la fonction $f(x)$ entraîne que $[f(\xi)]^2 \leqslant 0$. Il s'ensuit que $f(\xi) = 0$, c'est-à-dire que ξ est une racine de l'équation (1) et il est clair que

$$0 \leqslant \xi - a_n \leqslant \frac{1}{2^n}(b-a). \tag{4}$$

Si sur le segment $[a, b]$ les racines de l'équation (1) ne sont pas séparées, on peut utiliser ce procédé pour chercher l'une des racines de l'équation (1).

La méthode de bipartition est commode pour obtenir une estimation grossière d'une racine de l'équation donnée, le volume du calcul à effectuer marquant un net accroissement avec une précision plus élevée.

Constatons que la méthode de bipartition se réalise sans peine sur les calculateurs électroniques. Le programme de calcul est composé de façon que la machine fournit la valeur du deuxième membre de l'équation (1) au milieu de chacun des segments $[a_n, b_n]$ $(n = 1, 2, \ldots)$ et choisisse la moitié correspondante.

E x e m p l e. Améliorer par la méthode de bipartition la racine de l'équation

$$f(x) \equiv x^4 + 2x^3 - x - 1 = 0,$$

comprise dans le segment $[0, 1]$.

S o l u t i o n. On a successivement:

$$f(0) = -1;\ f(1) = 1;$$
$$f(0{,}5) = 0{,}06 + 0{,}25 - 0{,}5 - 1 = -1{,}19;$$
$$f(0{,}75) = 0{,}32 + 0{,}84 - 0{,}75 - 1 = -0{,}59;$$
$$f(0{,}875) = 0{,}59 + 1{,}34 - 0{,}88 - 1 = +0{,}05;$$
$$f(0{,}8125) = 0{,}436 + 1{,}072 - 0{,}812 - 1 = -0{,}304;$$
$$f(0{,}8438) = 0{,}507 + 1{,}202 - 0{,}844 - 1 = -0{,}135;$$
$$f(0{,}8594) = 0{,}546 + 1{,}270 - 0{,}859 - 1 = -0{,}043, \text{ etc.}$$

On peut poser

$$\xi = \frac{1}{2}(0{,}859 + 0{,}875) = 0{,}867.$$

§ 4. Méthode des parties proportionnelles

Indiquons (sous les hypothèses du § 3) une méthode de recherche plus rapide de la racine ξ de l'équation $f(x) = 0$, appartenant au segment considéré $[a, b]$ tel que $f(a) f(b) < 0$.

Soit, pour fixer les idées, $f(a) < 0$ et $f(b) > 0$. Alors, au lieu de diviser le segment $[a, b]$ en deux, il est plus logique de le diviser dans le rapport $-f(a) : f(b)$. On obtient ainsi la valeur approchée de la racine

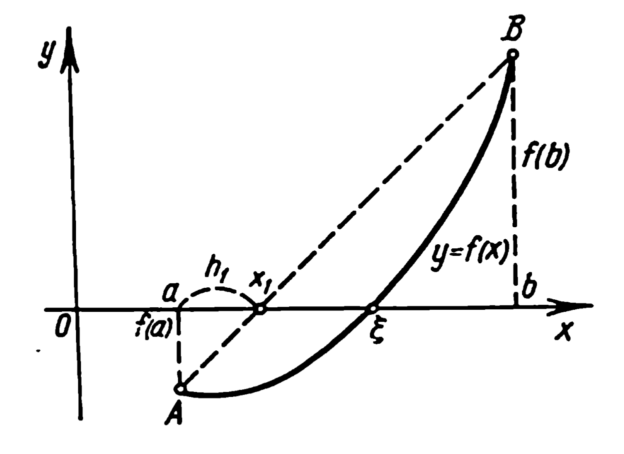


Fig. 15.

$$x_1 = a + h_1, \qquad (1)$$

où

$$h_1 = \frac{-f(a)}{-f(a)+f(b)}(b-a) = -\frac{f(a)}{f(b)-f(a)}(b-a). \qquad (2)$$

En appliquant ensuite ce procédé à celui des segments $[a, x_1]$ ou $[x_1, b]$ aux extrémités duquel les signes de la fonction $f(x)$ sont contraires, on obtient la deuxième approximation x_2 de la racine, etc.

Géométriquement, la méthode des parties proportionnelles est équivalente au remplacement de la courbe $y = f(x)$ par une corde menée par les points $A\,[a, f(a)]$ et $B\,[b, f(b)]$ (fig. 15). En effet, l'équation de la corde AB s'écrit

$$\frac{x-a}{b-a} = \frac{y-f(a)}{f(b)-f(a)}.$$

En posant $x = x_1$ et $y = 0$, on tire

$$x_1 = a - \frac{f(a)}{f(b)-f(a)}(b-a). \qquad (1')$$

La formule (1') est parfaitement équivalente aux formules (1) et (2).

Pour démontrer la convergence du processus, supposons que la racine est séparée et que sur le segment $[a, b]$ le signe de la dérivée seconde $f''(x)$ est constant.

Soit, pour fixer les idées, $f''(x) > 0$ avec $a \leqslant x \leqslant b$ (pour ramener $f''(x) < 0$ à notre cas l'équation doit être mise sous la forme $-f(x) = 0$). La courbe $y = f(x)$ sera alors convexe vers le bas et, par conséquent, elle se situera au-dessous de sa corde AB. Deux cas sont alors possibles: 1) $f(a) > 0$ (fig. 16) et 2) $f(a) < 0$ (fig. 17).

Dans le premier cas, l'extrémité a est fixe et les approximations successives: $x_0 = b$;

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n)-f(a)}(x_n - a) \quad (n = 0, 1, 2, \ldots) \qquad (3)$$

forment une suite décroissante bornée, en outre,

$$a < \xi < \ldots < x_{n+1} < x_n < \ldots < x_1 < x_0.$$

Dans le deuxième cas, c'est l'extrémité b qui est fixe et les approximations successives: $x_0 = a$;

$$x_{n+1} = x_n - \frac{f(x_n)}{f(b) - f(x_n)} (b - x_n) \tag{4}$$

forment une suite croissante bornée, de plus,

$$x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} < \ldots < \xi < b.$$

La généralisation de ces résultats conduit à la conclusion suivante: 1) l'extrémité fixe est celle dont le signe de la fonction $f(x)$

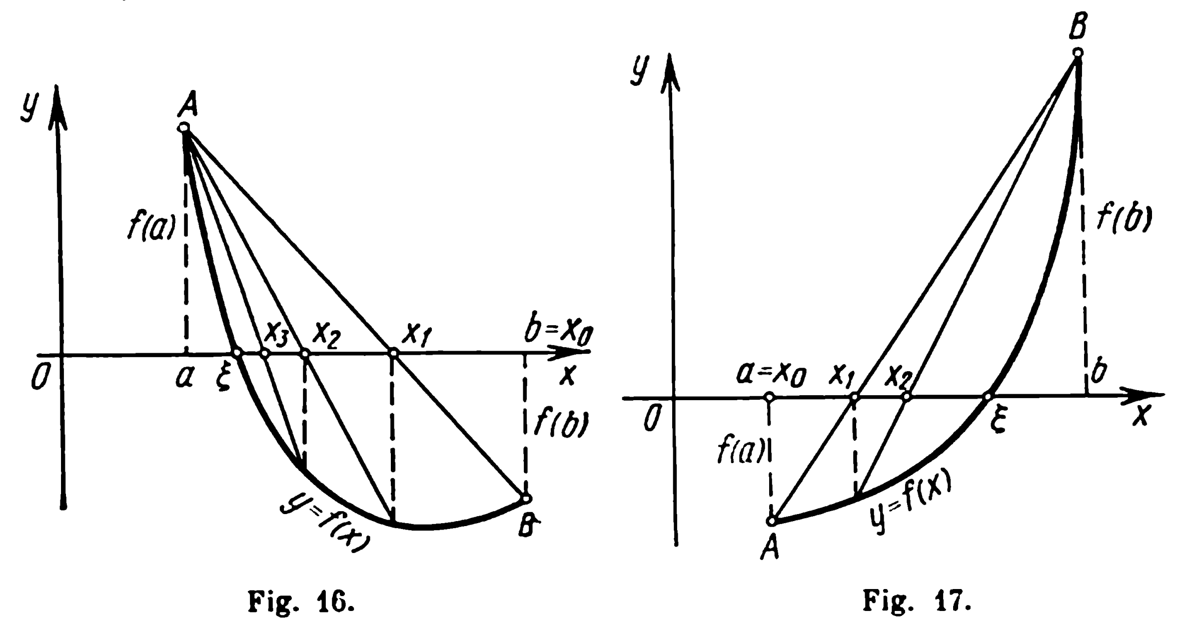


Fig. 16. Fig. 17.

coïncide avec le signe de sa dérivée seconde $f''(x)$; 2) les approximations successives x_n reposent du côté de la racine ξ où la fonction $f(x)$ a un signe opposé à celui de sa dérivée seconde $f''(x)$. Dans les deux cas chaque approximation successive x_{n+1} est plus proche de la racine ξ que l'approximation précédente x_n. Soit

$$\bar{\xi} = \lim_{n \to \infty} x_n \qquad (a < \bar{\xi} < b)$$

(la limite existe du fait que la suite $\{x_n\}$ est bornée et monotone). En passant dans l'égalité (3) à la limite, on a pour le premier cas:

$$\bar{\xi} = \bar{\xi} - \frac{f(\bar{\xi})}{f(\bar{\xi}) - f(a)} (\bar{\xi} - a);$$

on en tire $f(\bar{\xi}) = 0$. Par hypothèse, l'équation $f(x) = 0$ admet dans l'intervalle (a, b) une seule racine ξ, donc $\bar{\xi} = \xi$, ce qu'il fallait démontrer.

D'une façon analogue, en passant à la limite dans l'égalité (4), on montre pour le deuxième cas que $\bar{\xi} = \xi$.

Pour évaluer la précision de l'approximation on peut utiliser la formule (5) du § 1

$$|x_n - \xi| \leqslant \frac{|f(x_n)|}{m_1},$$

où $|f'(x)| \geqslant m_1$ avec $a \leqslant x \leqslant b$.

Voici encore une formule qui permet d'évaluer l'erreur absolue de la valeur approchée x_n si l'on connaît deux approximations successives x_{n-1} et x_n.

Supposons que la dérivée $f'(x)$ soit continue et de signe constant sur le segment $[a, b]$ qui contient toutes les approximations et de plus

$$0 < m_1 \leqslant |f'(x)| \leqslant M_1 < +\infty. \tag{5}$$

Posons, pour fixer les idées, que les approximations successives x_n de la racine exacte ξ sont élaborées d'après la formule (3) (l'exploration de la formule (4) est analogue)

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f(x_{n-1}) - f(a)}(x_{n-1} - a)$$

($n = 1, 2, \ldots$), où l'extrémité a est fixe. Il s'ensuit en tenant compte de $f(\xi) = 0$ que

$$f(\xi) - f(x_{n-1}) = \frac{f(x_{n-1}) - f(a)}{x_{n-1} - a}(x_n - x_{n-1}).$$

En appliquant le théorème de Lagrange des accroissements finis, il vient :

$$(\xi - x_{n-1}) f'(\xi_{n-1}) = (x_n - x_{n-1}) f'(\bar{x}_{n-1}),$$

où $\xi_{n-1} \in (x_{n-1}, \xi)$ et $\bar{x}_{n-1} \in (a, x_{n-1})$. Par conséquent,

$$|\xi - x_n| = \frac{|f'(\bar{x}_{n-1}) - f'(\xi_{n-1})|}{|f'(\xi_{n-1})|}|x_n - x_{n-1}|. \tag{6}$$

Etant donné que sur le segment $[a, b]$ $f'(x)$ garde le signe constant et, en outre, $\bar{x}_{n-1} \in [a, b]$ et $\xi_{n-1} \in [a, b]$, on a bien

$$|f'(\bar{x}_{n-1}) - f'(\xi_{n-1})| \leqslant M_1 - m_1.$$

La formule (6) amène donc :

$$|\xi - x_n| \leqslant \frac{M_1 - m_1}{m_1}|x_n - x_{n-1}|, \tag{7}$$

où on peut adopter comme m_1 et M_1 respectivement les valeurs minimale et maximale du module de la dérivée $f'(x)$ sur le segment $[a, b]$. Si le segment $[a, b]$ est tellement étroit qu'il donne lieu à l'inégalité

$$M_1 \leqslant 2m_1,$$

la formule (7) entraîne:

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|.$$

Ainsi, dans ce cas, dès que nous découvrons que

$$|x_n - x_{n-1}| < \varepsilon,$$

où ε est la borne d'erreur absolue donnée, nous aurons sûrement

$$|\xi - x_n| < \varepsilon.$$

Exemple. Trouver la racine positive de l'équation

$$f(x) \equiv x^3 - 0,2x^2 - 0,2x - 1,2 = 0$$

à 0,002 près.

Solution. Tout d'abord séparons la racine. Puisque

$$f(1) = -0,6 < 0 \text{ et } f(2) = 5,6 > 0,$$

la racine cherchée ξ appartient à l'intervalle (1, 2). L'intervalle obtenu est grand et nous le partagerons en deux. Etant donné que

$$f(1,5) = 1,425, \quad \text{on a} \quad 1 < \xi < 1,5.$$

L'application successive des formules (1) et (2) conduit à

$$x_1 = 1 + \frac{0,6}{1,425+0,6}(1,5-1) = 1 + 0,15 = 1,15;$$

$$f(x_1) = -0,173;$$

$$x_2 = 1,15 + \frac{0,173}{1,425+0,173}(1,5-1,15) = 1,15 + 0,040 = 1,190;$$

$$f(x_2) = -0,036;$$

$$x_3 = 1,190 + \frac{0,036}{1,425+0,036}(1,5-1,190) = 1,190 + 0,008 = 1,198;$$

$$f(x_3) = -0,0072.$$

Comme $f'(x) = 3x^2 - 0,4x - 0,2$ et avec $x_3 < x < 1,5$ on a

$$f'(x) \geqslant 3 \cdot 1,198^2 - 0,4 \cdot 1,5 - 0,2 = 3 \cdot 1,43 - 0,8 = 3,49,$$

on peut poser

$$0 < \xi - x_3 < \frac{0,0072}{3,49} \approx 0,002.$$

Ainsi, $\xi = 1,198 + 0,002\theta$ où $0 < \theta \leqslant 1$.

Constatons que la racine exacte de l'équation (5) est $\xi = 1,2$.

§ 5. Méthode de Newton

Soit la racine ξ de l'équation

$$f(x) = 0 \tag{1}$$

séparée sur le segment $[a, b]$; de plus $f'(x)$ et $f''(x)$ sont continues et gardent des signes constants pour $a \leqslant x \leqslant b$. Après le calcul d'une

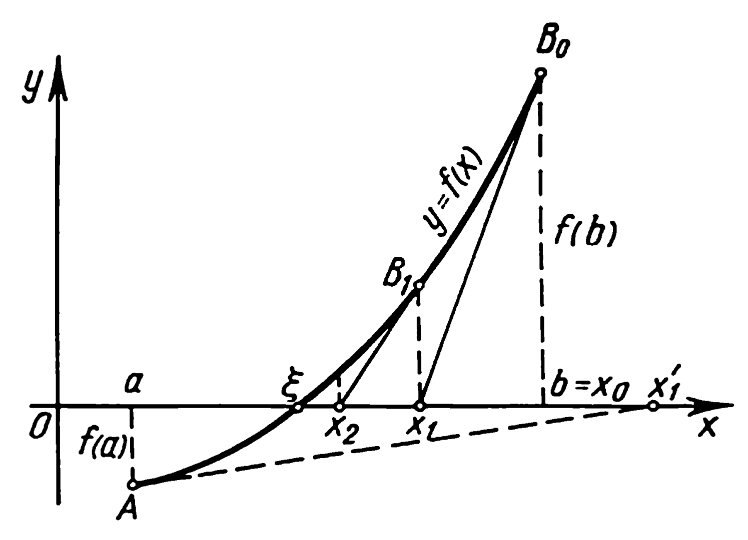


Fig. 18.

n-ième valeur approchée de la racine $x_n \approx \xi$ $(a \leqslant x_n \leqslant b)$ nous pouvons améliorer sa précision de la façon suivante en recourant à la *méthode de Newton*. Posons

$$\xi = x_n + h_n, \tag{2}$$

où h_n est une petite grandeur. D'où, en appliquant la formule de Taylor,

$$0 = f(x_n + h_n) \approx f(x_n) + h_n f'(x_n).$$

Par conséquent,

$$h_n = -\frac{f(x_n)}{f'(x_n)}.$$

Après avoir introduit cette correction dans la formule (2) on trouve l'approximation successive (dans l'ordre considéré) de la racine

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (n = 0, 1, 2, \ldots). \tag{3}$$

Géométriquement la méthode de Newton est équivalente au remplacement d'un petit arc de la courbe $y = f(x)$ par la tangente menée par un certain point de cette courbe. En effet, posons pour fixer les idées, que $f''(x) > 0$ avec $a \leqslant x \leqslant b$ et $f(b) > 0$ (fig. 18).

Choisissons, par exemple, $x_0 = b$ tel que $f(x_0) f''(x_0) > 0$. Menons par le point B_0 $[x_0, f(x_0)]$ la tangente à la courbe $y = f(x)$. Prenons l'abscisse du point d'intersection de cette tangente avec l'axe Ox comme première approximation x_1 de la racine ξ. Menons encore une fois par le point B_1 $[x_1, f(x_1)]$ la tangente dont l'abscisse du point d'intersection donnera la deuxième approximation x_2 de la racine ξ, etc. (fig. 18). Il est clair que l'équation de la tangente en B_n $[x_n, f(x_n)]$ $(n = 0, 1, 2, \ldots)$ s'écrit

$$y - f(x_n) = f'(x_n)(x - x_n).$$

En posant $y = 0$, $x = x_{n+1}$, on obtient la formule (3)

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Constatons que si dans le cas considéré on adopte $x_0 = a$ et, par conséquent, $f(x_0) f''(x_0) < 0$, en menant la tangente à la courbe $y = f(x)$ par A $[a, f(a)]$, on obtiendrait le point x_1 (fig. 18) situé hors du segment $[a, b]$, c'est-à-dire ce choix de la valeur initiale rend la méthode de Newton peu avantageuse. Ainsi, dans le cas considéré, une « bonne » approximation initiale x_0 est celle qui vérifie l'inégalité

$$f(x_0) f''(x_0) > 0. \tag{4}$$

Montrons que cette règle est générale.

Théorème. *Si $f(a) f(b) < 0$, si en outre $f'(x)$ et $f''(x)$ sont non nuls et gardent des signes constants pour $a \leqslant x \leqslant b$, la racine unique ξ de l'équation* (1) *peut être calculée à l'aide de la méthode de Newton (formule* (3)) *avec la précision aussi grande que l'on veut, en partant de l'approximation initiale $x_0 \in [a, b]$ qui satisfait à l'inégalité* (4).

Démonstration. Soit, par exemple, $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) > 0$ avec $a \leqslant x \leqslant b$ (la discussion des autres cas est analogue). D'après l'inégalité (4) on a $f(x_0) > 0$ (on peut poser, par exemple, $x_0 = b$).

Montrons par récurrence que toute approximation $x_n > \xi$ $(n = 0, 1, 2, \ldots)$ et, par conséquent, $f(x_n) > 0$. En effet, tout d'abord $x_0 > \xi$.

Soit maintenant $x_n > \xi$. Posons

$$\xi = x_n + (\xi - x_n).$$

En appliquant la formule de Taylor, on a:

$$0 = f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2} f''(c_n)(\xi - x_n)^2, \tag{5}$$

où $\xi < c_n < x_n$.

Etant donné que $f''(x) > 0$, il vient:

$$f(x_n) + f'(x_n)(\xi - x_n) < 0$$

et, par conséquent,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > \xi,$$

ce qu'il fallait démontrer.

En tenant compte des signes de $f(x_n)$ et de $f'(x_n)$, la formule (3) implique $x_{n+1} < x_n$ ($n = 0, 1, 2, \ldots$), c'est-à-dire les approximations successives $x_0, x_1, \ldots, x_n, \ldots$ forment une suite décroissante bornée. Il existe donc une limite $\bar{\xi} = \lim_{n\to\infty} x_n$.

En passant à la limite dans l'égalité (3) on a:

$$\bar{\xi} = \bar{\xi} - \frac{f(\bar{\xi})}{f'(\bar{\xi})}$$

ou $f(\bar{\xi}) = 0$. On en tire $\bar{\xi} = \xi$, et le théorème est ainsi démontré.

C'est pourquoi en appliquant la méthode de Newton il faut se guider sur la règle suivante: *choisir comme point initial x_0 celle des extrémités de l'intervalle (a, b) à laquelle correspond l'ordonnée de même signe que $f''(x)$.*

R e m a r q u e 1. Si 1) la fonction $f(x)$ est définie et continue pour $-\infty < x < +\infty$; 2) $f(a)f(b) < 0$; 3) $f'(x) \neq 0$ pour $a \leqslant x \leqslant b$; 4) $f''(x)$ existe partout et garde le signe constant, alors, en appliquant la méthode de Newton pour chercher la racine de l'équation $f(x) = 0$ comprise dans l'intervalle (a, b), on peut prendre comme approximation initiale x_0 une valeur quelconque $c \in [a, b]$. En particulier, on peut prendre $x_0 = a$ ou $x_0 = b$.

En effet, soit, par exemple, $f'(x) > 0$ pour $a \leqslant x \leqslant b$, $f''(x) > 0$ et $x_0 = c$, où $a \leqslant c \leqslant b$.

Si $f(c) = 0$, la racine $\xi = c$ et le problème est ainsi résolu.

Si $f(c) > 0$, le raisonnement ci-dessus se trouve justifié et le processus de Newton à valeur initiale c converge vers la racine $\xi \in (a, b)$.

Enfin, si $f(c) < 0$, on tombe sur la valeur

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = c - \frac{f(c)}{f'(c)} > c.$$

En appliquant la formule de Taylor, on a:

$$f(x_1) = f(c) - \frac{f(c)}{f'(c)} f'(c) + \frac{1}{2}\left[\frac{f(c)}{f'(c)}\right]^2 f''(\bar{c}) = \frac{1}{2}\left[\frac{f(c)}{f'(c)}\right]^2 f''(\bar{c}) > 0,$$

où $\bar{c}$ est une certaine valeur intermédiaire entre c et x_1.

Ainsi,

$$f(x_1)\, f''(x_1) > 0.$$

Par ailleurs, la condition $f''(x) > 0$ entraîne que $f'(x)$ est une fonction croissante et, par conséquent, $f'(x) > f'(a) > 0$ avec $x > a$. On peut donc prendre x_1 comme valeur initiale du processus de Newton qui converge vers une certaine racine $\bar{\xi}$ de la fonction $f(x)$ telle que $\bar{\xi} > c \geqslant a$. Puisque la positivité de la dérivée $f'(x)$ avec $x > a$ implique que la fonction $f(x)$ possède une seule racine dans l'intervalle $(a, +\infty)$, il vient

$$\bar{\xi} = \xi \in (a, b).$$

Une analyse analogue peut être appliquée à d'autres combinaisons de signes des dérivées $f'(x)$ et $f''(x)$.

R e m a r q u e 2. La formule (3) montre que plus la valeur numérique de la dérivée $f'(x)$ est grande dans le voisinage de la racine considérée, plus la correction qu'il faut ajouter à la n-ième approximation pour obtenir la $(n + 1)$-ième approximation est petite. Il s'ensuit que la méthode de Newton est surtout commode lorsque dans le voisinage de la racine considérée la pente du graphe de la fonction est importante. Mais si la valeur numérique de la dérivée $f'(x)$ dans le voisinage de la racine est faible, les corrections seront grandes et le calcul de la racine d'après cette méthode peut prendre beaucoup de temps et devenir même impossible. Par conséquent, si la courbe $y = f(x)$ est presque horizontale dans le voisinage du point d'intersection avec l'axe Ox, l'utilisation de la méthode de Newton pour la résolution de l'équation $f(x) = 0$ n'est pas recommandée.

Pour évaluer l'erreur de la n-ième approximation x_n, on peut faire appel à la formule générale (5) du § 1

$$|\xi - x_n| \leqslant \frac{|f(x_n)|}{m_1}, \tag{6}$$

où m_1 est la valeur minimale de $|f'(x)|$ sur le segment $[a, b]$.

Déduisons encore une formule pour l'estimation de la précision de l'approximation x_n. En appliquant la formule de Taylor on a:

$$f(x_n) = f[x_{n-1} + (x_n - x_{n-1})] =$$
$$= f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \frac{1}{2} f''(\xi_{n-1})(x_n - x_{n-1})^2, \tag{7}$$

où $\xi_{n-1} \in (x_{n-1}, x_n)$. Puisque la définition de l'approximation x_n donne

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0,$$

(7) conduit à

$$|f(x_n)| \leqslant \frac{1}{2} M_2 (x_n - x_{n-1})^2,$$

où M_2 est la valeur maximale de $|f''(x)|$ sur le segment $[a, b]$. Par conséquent, d'après la formule (6) on a finalement:

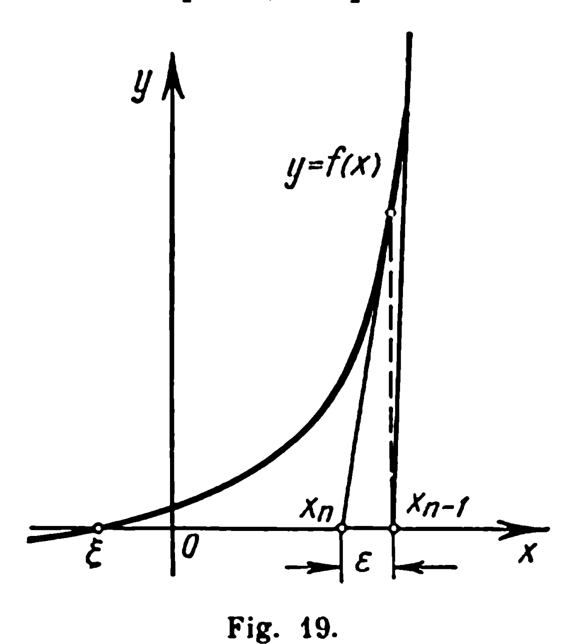


Fig. 19.

$$|\xi - x_n| \leqslant \frac{M_2}{2m_1}(x_n - x_{n-1})^2. \quad (8)$$

Si le processus de Newton converge, $x_n - x_{n-1} \to 0$ quand $n \to \infty$. C'est pourquoi avec $n \geqslant N$, on a:

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|,$$

quand N est suffisamment grand, ce qui signifie qu'à partir d'une certaine approximation les premiers chiffres « *stabilisés* » des approximations x_{n-1} et x_n sont exacts.

Constatons que dans le cas général la coïncidence à ε près de deux approximations successives x_{n-1} et x_n n'assure nullement la coïncidence avec la même précision de la valeur x_n et de la racine exacte ξ (fig. 19).

Etablissons également la formule qui associe les erreurs absolues de deux approximations successives x_n et x_{n+1}. La formule (5) entraîne

$$\xi = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2} \cdot \frac{f''(c_n)}{f'(c_n)}(\xi - x_n)^2,$$

où $c_n \in (x_n, \xi)$. D'où, en tenant compte de la formule (3), on a

$$\xi - x_{n+1} = -\frac{1}{2} \cdot \frac{f''(c_n)}{f'(x_n)}(\xi - x_n)^2$$

et, par conséquent,

$$|\xi - x_{n+1}| \leqslant \frac{M_2}{2m_1}(\xi - x_n)^2. \quad (9)$$

La formule (9) assure une convergence rapide du processus de Newton si l'approximation initiale x_0 est telle que

$$\frac{M_2}{2m_1}|\xi - x_0| \leqslant q < 1.$$

En particulier, si

$$\mu = \frac{M_2}{2m_1} \leqslant 1 \quad \text{et} \quad |\xi - x_n| < 10^{-m},$$

la formule (9) conduit à

$$|\xi - x_{n+1}| < 10^{-2m},$$

c'est-à-dire, dans ce cas, si l'approximation x_n compte m décimales exactes, l'approximation suivante x_{n+1} comptera au moins $2m$ décimales exactes; autrement dit, si $\mu \leqslant 1$, à l'aide de la méthode de Newton le nombre de décimales exactes de la racine cherchée ξ est doublé à chaque étape.

Exemple 1. Calculer par la méthode de Newton la racine négative de l'équation $f(x) \equiv \equiv x^4 - 3x^2 + 75x - 10\,000 = 0$ avec cinq chiffres exacts.

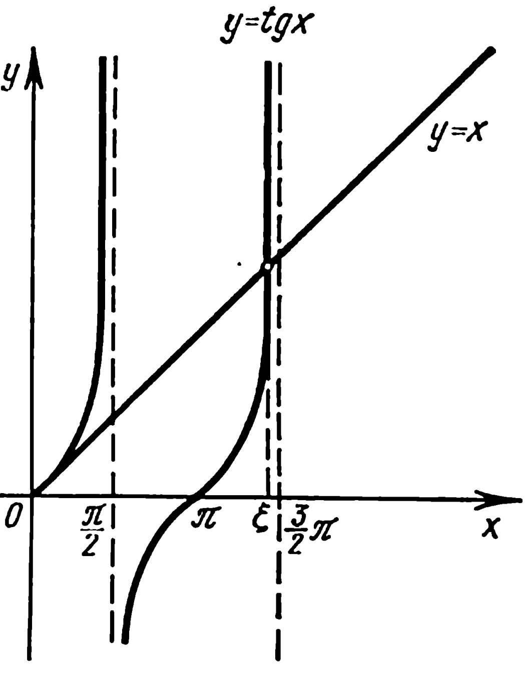


Fig. 20.

Solution. En posant dans le premier membre de l'équation $x = 0, -10, -100, \ldots$, on a $f(0) = -10000, f(-10) = -1050$, $f(-100) \approx + 10^8$.

Par conséquent, la racine cherchée ξ repose dans l'intervalle $-100 < \xi < -10$. Réduisons l'intervalle obtenu. Etant donné que $f(-11) = 3453$, on a $-11 < \xi < -10$. Dans ce dernier intervalle $f'(x) < 0$ et $f''(x) > 0$. Puisque $f(-11) > 0$ et $f''(-11) > 0$, nous pouvons adopter comme approximation initiale $x_0 = -11$. Calculons les approximations successives x_n $(n = 1, 2, \ldots)$ d'après le schéma suivant:

n	x_n	$f(x_n)$	$f'(x_n)$	$h_n = -\frac{f(x_n)}{f'(x_n)}$
0	−11	3453	−5183	0,7
1	−10,3	134,3	−4234	0,03
2	−10,27	37,8	−4196	0,009
3	−10,261	0,2	—	—

En optant pour $n = 3$, vérifions le signe de la valeur $f(x_n + 0{,}001) = f(-10{,}260)$. Puisque $f(-10{,}260) < 0$, on a $-10{,}261 < \xi < -10{,}260$ et n'importe quel de ces nombres donne l'approximation cherchée.

Exemple 2. Trouver par la méthode de Newton la racine positive minimale de l'équation $\operatorname{tg} x = x$ à 0,0001 près.

Solution. Construisons les courbes $y = \operatorname{tg} x$ et $y = x$ (fig. 20) pour constater que la racine cherchée ξ appartient à l'inter-

valle $\pi < \xi < \frac{3\pi}{2}$. En mettant l'équation sous la forme

$$f(x) \equiv \sin x - x \cos x = 0,$$

on a:

$$f'(x) = x \sin x;$$

$$f''(x) = \sin x + x \cos x.$$

Il s'ensuit que $f'(x) < 0$ et $f''(x) < 0$ pour $\pi < x < \frac{3\pi}{2}$. Comme $f\left(\frac{3\pi}{2}\right) = -1$, on peut prendre comme approximation initiale $x_0 = \frac{3\pi}{2}$. Le calcul se fait d'après le schéma suivant:

n	x_n	$f(x_n)$	$f'(x_n)$	$h_n = -\frac{f(x_n)}{f'(x_n)}$
0	$\frac{3\pi}{2}$ =4,71239 (270°)	−1	−4,712	−0,212 (≈ −12°10′)
1	4,50004 (257°50′)	−0,0291	−4,399	−0,0066 (≈ −22′42″)
2	4,49343 (257°27′16″)	−0,00003	—	—

Pour évaluer l'erreur de la valeur approchée x_n constatons qu'en vertu de la négativité de la dérivée seconde $f''(x)$ les approximations successives x_n $(n = 0, 1, 2, \ldots)$ sont décroissantes; en outre, $f(x_n) < 0$. C'est pourquoi on peut poser $\bar{x}_n < \xi < x_n$, où $\bar{x}_n$ est un nombre de l'intervalle $\left(\pi, \frac{3\pi}{2}\right)$ tel que $f(\bar{x}_n) > 0$. La valeur $\bar{x}_n$ s'obtient sans peine par sélection *. Ainsi, avec $n = 2$ et en posant approximativement

$$\bar{x}_2 = 4{,}49340 = \text{arc } 257°27'12'',$$

on a

$$f(\bar{x}_2) = \sin 257°27'12'' - 4{,}49340 \cdot \cos 257°27'12'' = -0{,}97612 + $$
$$+ 4{,}49340 \cdot 0{,}21724 = -0{,}97612 + 0{,}97614 = +0{,}00002.$$

Par conséquent, le choix de $\bar{x}_2$ est correct et donc

$$4{,}49340 < \xi < 4{,}49343.$$

On peut poser

$$\xi = 4{,}4934,$$

où tous les chiffres sont exacts.

---

* On pourrait prendre, certes, $\bar{x}_n = \pi$, mais un tel choix est défavorable du fait que $f'(\pi) = 0$.

L'erreur de x_2 peut être évaluée d'une façon plus précise. Puisque avec $x \in [\bar{x}_2, x_2]$ la dérivée $f'(x)$ décroît et $f'(x) < 0$, on a

$$m_1 = \min |f'(x)| = |f'(\bar{x}_2)|.$$

Il en résulte:

$$m_1 = 4{,}49340 \cdot 0{,}97612 > 4$$

et, par conséquent,

$$|\xi - x_2| \leqslant \frac{|f(x_2)|}{4} = \frac{0{,}00003}{4} < 10^{-5}.$$

Ainsi

$$\xi = 4{,}49343 - 0{,}00001\theta,$$

où $0 < \theta < 1$.

Exemple 3. Considérons l'équation

$$f(x) = 0, \qquad (10)$$

où $f''(x)$ est continue et de signe constant pour $-\infty < x < +\infty$. En vertu du théorème de Rolle l'équation (10) ne peut avoir plus de deux racines réelles. Voici deux cas importants pour la pratique.

I. Soit

$$f(x_0)\, f'(x_0) < 0,$$
$$f(x_0)\, f''(x) < 0$$

(fig. 21).

L'équation (10) admet alors une racine unique ξ dans l'intervalle (x_0, x_1), où

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

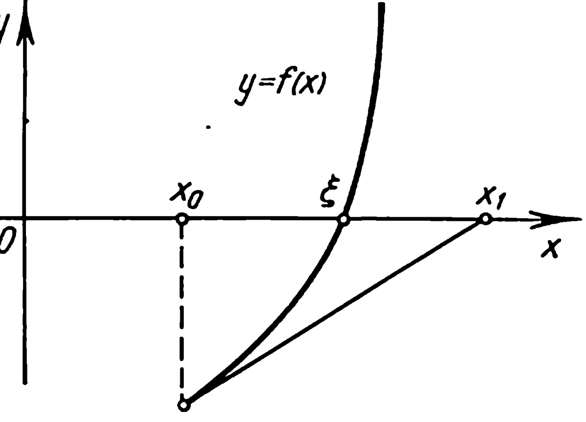


Fig. 21.

La racine ξ peut être calculée avec la précision imposée à l'aide de la méthode de Newton.

II. Soit

$$f'(x_0) = 0, \qquad f(x_0)\, f''(x) < 0.$$

L'équation (10) a alors deux racines ξ et ξ' dans l'intervalle $(-\infty, +\infty)$ (fig. 22).

La transformation du premier membre de l'équation (10) d'après la formule de Taylor donne approximativement:

$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 = 0$$

ou

$$f(x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 = 0.$$

On en tire pour les racines ξ et ξ' les approximations initiales

$$x_1 = x_0 - \sqrt{-\frac{2f(x_0)}{f''(x_0)}}$$

et

$$x_1' = x_0 + \sqrt{-\frac{2f(x_0)}{f''(x_0)}},$$

qui sont les abscisses des points d'intersection de la parabole

$$Y = f(x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

avec l'axe Ox (fig. 23). L'amélioration de la précision des racines peut s'obtenir par la méthode de Newton usuelle.

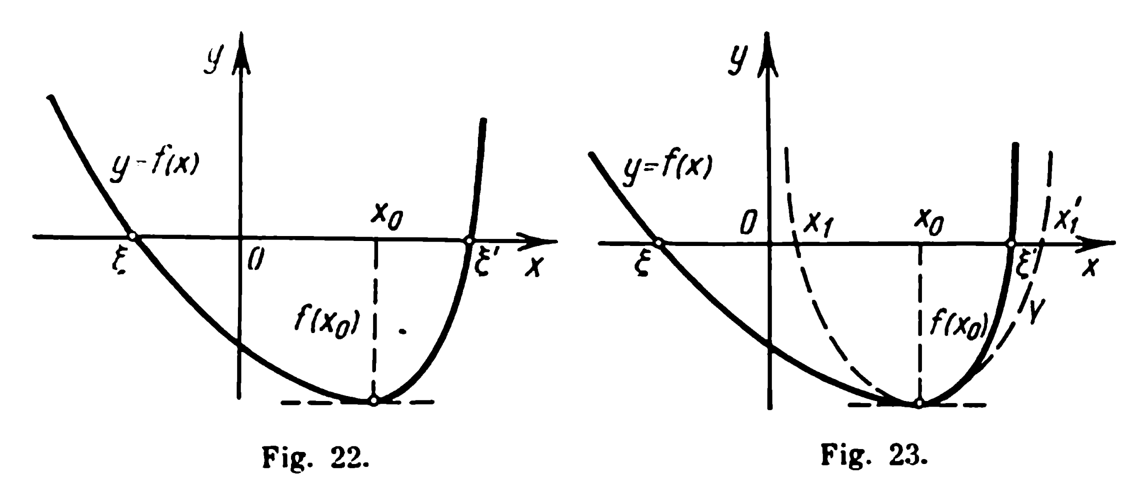


Fig. 22. Fig. 23.

Le sens géométrique des affirmations I et II est évident. Nous laissons au lecteur le soin de réaliser leur démonstration rigoureuse.

§ 6. Méthode de Newton modifiée

Si la dérivée $f'(x)$ varie peu sur le segment $[a, b]$, on peut poser dans la formule (3) du paragraphe précédent:

$$f'(x_n) \approx f'(x_0). \quad (1)$$

Pour la racine ξ de l'équation $f(x) = 0$ on en tire les approximations successives

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \quad (n = 0, 1, \ldots). \quad (2)$$

L'interprétation géométrique de ce procédé est donnée par le remplacement des tangentes en $B_n\ [x_n, f(x_n)]$ par des droites parallèles à la tangente à la courbe $y = f(x)$ en son point fixe $B_0\ [x_0, f(x_0)]$ (fig. 24).

La formule (1) rend inutile le calcul repris chaque fois de la valeur de la dérivée $f'(x_n)$; c'est pourquoi cette formule est très

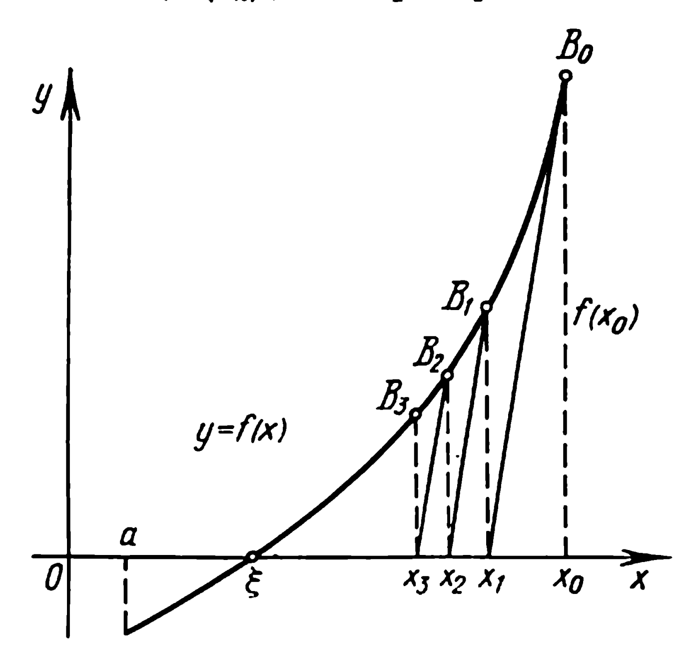


Fig. 24.

utile si $f'(x_n)$ est compliquée. On peut montrer que sous l'hypothèse de la permanence des signes des dérivées $f'(x)$ et $f''(x)$ les approximations successives (2) donnent un processus convergent.

§ 7. Méthode combinée

Soit $f(a) f(b) < 0$ alors que $f'(x)$ et $f''(x)$ gardent les signes constants sur le segment $[a, b]$. En combinant la méthode des parties proportionnelles à la méthode de Newton on obtient une méthode dont chaque étape permet de déterminer les valeurs par défaut et par excès de la racine exacte ξ de l'équation $f(x) = 0$.

Il en résulte, en particulier, que les chiffres communs pour x_n et $\bar{x}_n$ appartiennent nécessairement à la racine exacte ξ. Quatre cas peuvent se présenter théoriquement:

1) $f'(x) > 0$; $f''(x) > 0$ (fig. 25);

2) $f'(x) > 0$; $f''(x) < 0$ (fig. 26);

3) $f'(x) < 0$; $f''(x) > 0$ (fig. 27);

4) $f'(x) < 0$; $f''(x) < 0$ (fig. 28).

Nous nous bornerons à l'exploration du premier cas, l'étude des autres cas étant analogue; par ailleurs, le caractère des calculs se conçoit aisément à partir des dessins correspondants. Constatons que

tous ces cas peuvent être ramenés au premier si l'équation considérée $f(x) = 0$ est remplacée par des équations équivalentes: $-f(x) = 0$ ou $\pm f(-z) = 0$, où $z = -x$.

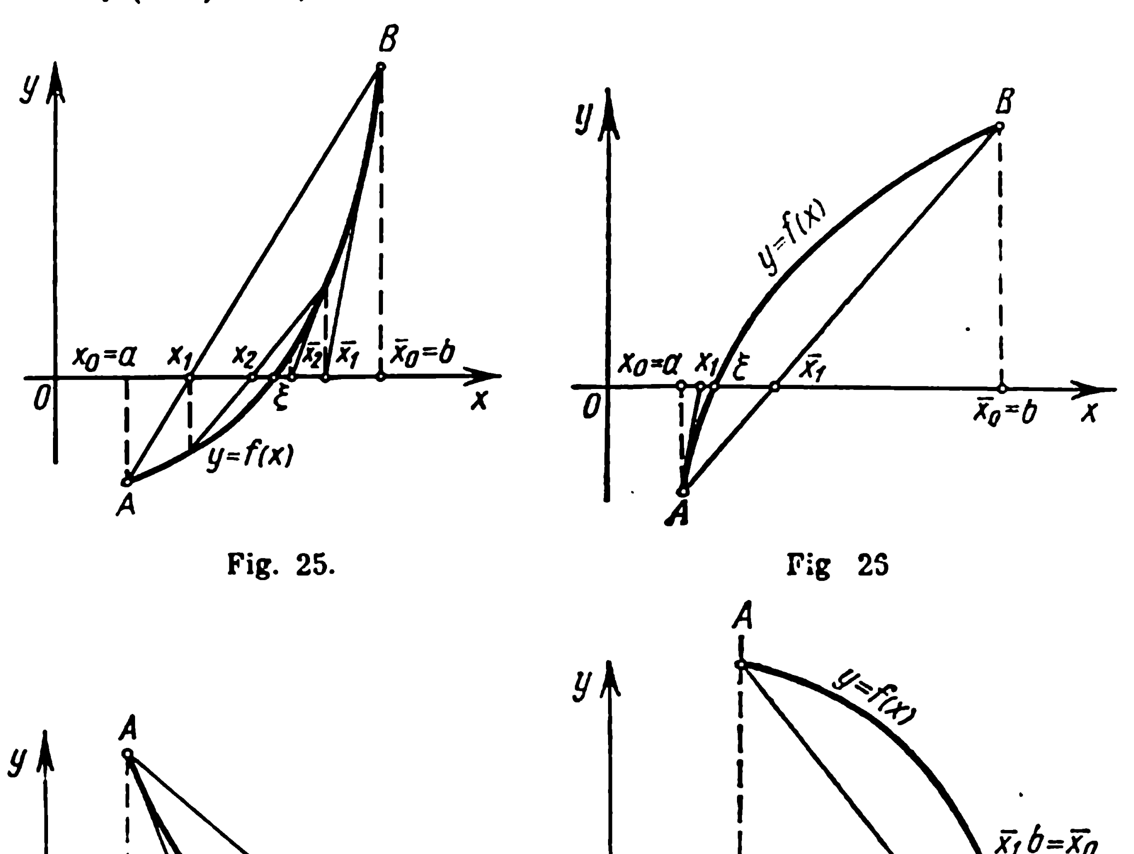


Fig. 25.

Fig 26

Fig. 27.

Fig. 28.

Ainsi, soit $f'(x) > 0$ et $f''(x) > 0$ pour $a \leqslant x \leqslant b$. Posons $x_0 = a$; $\bar{x}_0 = b$ et

$$x_{n+1} = x_n - \frac{f(x_n)}{f(\bar{x}_n) - f(x_n)} (\bar{x}_n - x_n)^* ; \qquad (1)$$

$$\bar{x}_{n+1} = \bar{x}_n - \frac{f(\bar{x}_n)}{f'(\bar{x}_n)} \qquad (n = 0, 1, 2, \ldots)^*. \qquad (1')$$

Les résultats des §§ 5 et 6 entraînent que

$$x_n < \xi < \bar{x}_n$$

---

* A chaque pas la méthode des parties proportionnelles est appliquée à un nouveau segment $[x_n, \bar{x}_n]$.

et

$$0 < \xi - x_n < \bar{x}_n - x_n. \qquad (2)$$

Si l'erreur absolue admissible de la racine approchée x_n est donnée à l'avance et est égale à ε, le processus de rapprochement s'arrête dès que l'on établit que $\bar{x}_n - x_n < \varepsilon$. Après la fin du processus le mieux est de prendre comme valeur de la racine ξ la moyenne arithmétique des dernières quantités obtenues:

$$\bar{\xi} = \frac{1}{2}(x_n + \bar{x}_n).$$

E x e m p l e. Calculer à 0,0005 près la racine positive de l'équation

$$f(x) \equiv x^5 - x - 0{,}2 = 0.$$

S o l u t i o n. Puisque $f(1) < 0$ et $f(1{,}1) > 0$, la racine appartient à l'intervalle $(1; 1{,}1)$. On a:

$$f'(x) = 5x^4 - 1 \quad \text{et} \quad f''(x) = 20x^3.$$

Dans l'intervalle choisi $f'(x) > 0$; $f''(x) > 0$, c'est-à-dire les signes des dérivées ne changent pas.

Appliquons la méthode combinée en posant $x_0 = 1$ et $\bar{x}_0 = 1{,}1$. Etant donné que

$$f(x_0) = f(1) = -0{,}2; \quad f(\bar{x}_0) = f(1{,}1) = 0{,}3105;$$

$$f'(\bar{x}_0) = f'(1{,}1) = 6{,}3205,$$

il ressort des formules (1) et (1')

$$x_1 = 1 + \frac{0{,}1 \cdot 0{,}2}{0{,}51051} \approx 1{,}039; \quad \bar{x}_1 = 1{,}1 - \frac{0{,}31051}{6{,}3205} \approx 1{,}051.$$

Vu que $\bar{x}_1 - x_1 = 0{,}012$, la précision est insuffisante. Cherchons le couple d'approximation suivant:

$$x_2 = 1{,}039 + \frac{0{,}012 \cdot 0{,}0282}{0{,}0595} \approx 1{,}04469; \quad \bar{x}_2 = 1{,}051 - \frac{0{,}0313}{5{,}1005} \approx 1{,}04487.$$

Ici $\bar{x}_2 - x_2 = 0{,}00018$, c'est-à-dire nous avons obtenu la précision imposée. On peut adopter

$$\bar{\xi} = \frac{1}{2}(1{,}04469 + 1{,}04487) = 1{,}04478 \approx 1{,}045$$

avec une erreur absolue inférieure à

$$\frac{1}{2} \cdot 0{,}00018 + 0{,}00022 = 0{,}00031 < \frac{1}{2} \cdot 10^{-3}.$$

§ 8. Méthode des approximations successives

Une des méthodes parmi les plus importantes de résolution numérique des équations est la *méthode des approximations successives* dite également méthode des itérations. Voici son principe. Soit l'équation

$$f(x) = 0, \qquad (1)$$

où $f(x)$ est une fonction continue; le problème consiste à déterminer ses racines réelles. Remplaçons l'équation (1) par une équation équivalente

$$x = \varphi(x). \qquad (2)$$

Sélectionnons par un moyen quelconque une valeur grossièrement approchée de la racine x_0 et portons-la dans le deuxième membre de l'équation (2). On obtient alors un certain nombre

$$x_1 = \varphi(x_0). \qquad (3)$$

Remplaçons maintenant dans le deuxième membre de l'égalité (3) x_0 par le nombre x_1 pour obtenir un nouveau nombre $x_2 = \varphi(x_1)$. En reprenant cette procédure, on aboutit à la suite des nombres

$$x_n = \varphi(x_{n-1}) \quad (n = 1, 2, \ldots). \qquad (4)$$

Si cette suite est convergente, c'est-à-dire s'il existe une limite $\xi = \lim\limits_{n\to\infty} x_n$, alors, en passant à la limite dans l'égalité (4) et en supposant la fonction $\varphi(x)$ continue, on tombe sur

$$\lim_{n\to\infty} x_n = \varphi(\lim_{n\to\infty} x_{n-1})$$

ou

$$\xi = \varphi(\xi). \qquad (5)$$

Ainsi, la limite ξ est une racine de l'équation (2) qui se calcule d'après la formule (4) avec la précision voulue.

Géométriquement cette méthode s'explique de la façon suivante. On construit dans le plan xOy les courbes des fonctions $y = x$ et $y = \varphi(x)$. Toute racine réelle ξ de l'équation (2) est l'abscisse d'un point d'intersection M de la courbe $y = \varphi(x)$ avec la droite $y = x$ (fig. 29).

En partant d'un certain point $A_0[x_0; \varphi(x_0)]$, on construit la ligne polygonale $A_0B_1A_1B_2A_2 \ldots$ (« échelonnée »), dont les éléments sont parallèles alternativement à l'axe Ox et à l'axe Oy, les sommets $A_0, A_1, A_2, \ldots$ reposent sur la courbe $y = \varphi(x)$, et les sommets $B_1, B_2, B_3, \ldots$ reposent sur la droite $y = x$. Les abscisses communes des points A_1 et B_1, A_2 et B_2, ... constituent respectivement les approximations successives $x_1, x_2, \ldots$ de la racine ξ.

La ligne polygonale $A_0B_1A_1B_2A_2$ . . . (fig. 30) peut avoir également une autre forme (« en spirale »). On conçoit aisément que la

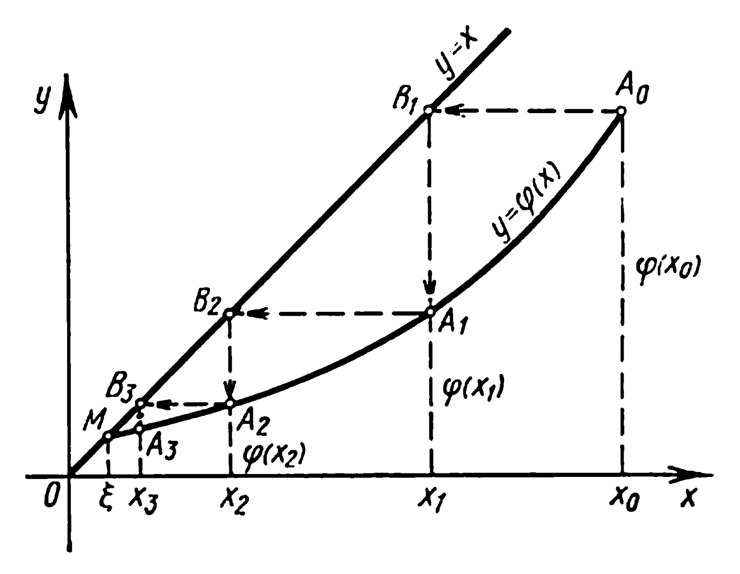


Fig. 29.

solution s'obtient sous forme d'une ligne « échelonnée » si la dérivée $\varphi'(x)$ est positive, et « en spirale » si $\varphi'(x)$ est négative.

Sur la figure 29 la pente de la courbe $y = \varphi(x)$ dans le voisinage de la racine ξ est faible, c'est-à-dire $|\varphi'(x)| < 1$ et le processus

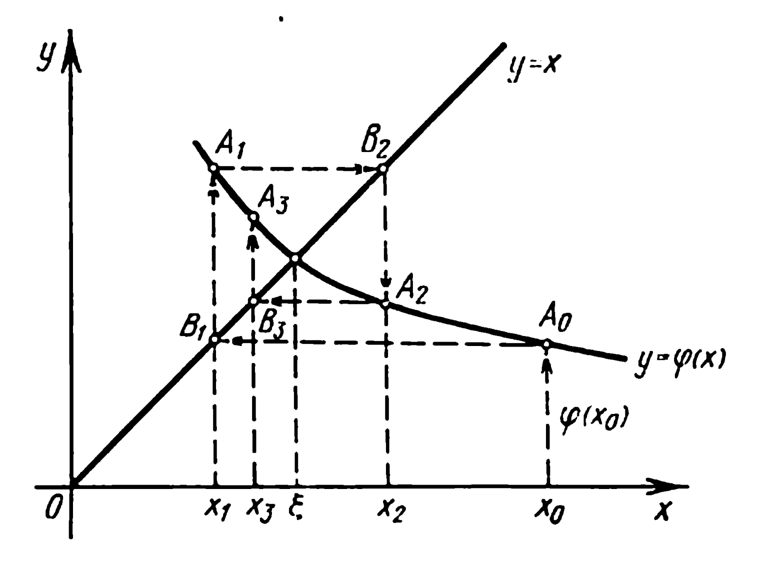


Fig. 30.

itératif converge. Toutefois, si l'on considère le cas où $|\varphi'(x)| > 1$, le processus itératif peut être divergent (fig. 31). Pour rendre possible l'application des approximations successives il faut donc définir les conditions suffisantes de convergence du processus itératif.

Théorème 1. *Soit la fonction* $\varphi(x)$ *définie et dérivable sur le segment* $[a, b]$ *avec toutes ses valeurs* $\varphi(x) \in [a, b]$.

S'il existe alors un nombre q tel que *

$$|\varphi'(x)| \leqslant q < 1 \tag{6}$$

pour $a < x < b$, 1) *le processus itératif*

$$x_n = \varphi(x_{n-1}) \qquad (n = 1, 2, \ldots) \tag{7}$$

converge indépendamment de la valeur initiale $x_0 \in [a, b]$; 2) *la valeur limite*

$$\xi = \lim_{n\to\infty} x_n$$

est l'unique racine de l'équation

$$x = \varphi(x) \tag{8}$$

sur le segment $[a, b]$.

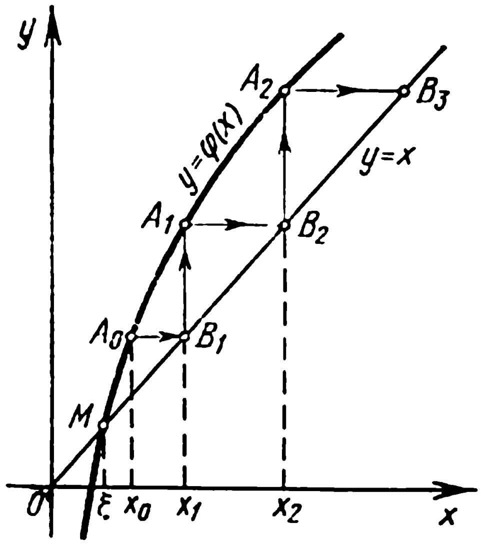


Fig. 31.

Démonstration. Considérons deux approximations successives

$$x_n = \varphi(x_{n-1}) \quad \text{et} \quad x_{n+1} = \varphi(x_n)$$

(qui en vertu des conditions du théorème ont bien un sens). On en tire

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1}).$$

En appliquant le théorème de Lagrange, on a

$$x_{n+1} - x_n = (x_n - x_{n-1})\, \varphi'(\bar{x}_n),$$

où $\bar{x}_n \in (x_{n-1}, x_n)$. Par conséquent, la condition (6) amène

$$|x_{n+1} - x_n| \leqslant q\, |x_n - x_{n-1}|. \tag{9}$$

Par suite, en donnant à n les valeurs 1, 2, 3, . . ., on déduit successivement

$$|x_2 - x_1| \leqslant q\, |x_1 - x_0|;$$

$$|x_3 - x_2| \leqslant q\, |x_2 - x_1| \leqslant q^2\, |x_1 - x_0|;$$

. . . .

$$|x_{n+1} - x_n| \leqslant q^n\, |x_1 - x_0|. \tag{10}$$

---

* On peut prendre comme nombre q la plus petite valeur ou la borne inférieure du module de la dérivée $|\varphi'(x)|$ pour $a \leqslant x \leqslant b$.

Considérons la série

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}) + \ldots, \quad (11)$$

telle que nos approximations successives x_n soient ses $(n+1)$-ièmes sommes partielles, c'est-à-dire

$$x_n = S_{n+1}.$$

Par suite de l'inégalité (10) les termes de la série (11) sont en valeur absolue inférieurs aux termes correspondants de la progression géométrique à raison $q < 1$, c'est pourquoi la série (11) converge et, de plus, d'une façon absolue. Par conséquent, il existe une limite

$$\lim_{n\to\infty} S_{n+1} = \lim_{n\to\infty} x_n = \xi,$$

en outre, on a bien $\xi \in [a, b]$.

La fonction $\varphi(x)$ étant continue, le passage à la limite dans l'égalité (7) conduit à

$$\xi = \varphi(\xi). \quad (12)$$

ξ est donc la racine de l'équation (8) qui n'a pas d'autres racines sur le segment $[a, b]$. En effet, si

$$\bar{\xi} = \varphi(\bar{\xi}), \quad (13)$$

les égalités (12) et (13) amènent

$$\bar{\xi} - \xi = \varphi(\bar{\xi}) - \varphi(\xi)$$

et, par conséquent,

$$(\bar{\xi} - \xi)[1 - \varphi'(c)] = 0, \quad (14)$$

où $c \in [\xi, \bar{\xi}]$. L'expression entre crochets de l'égalité (14) n'étant pas nulle, $\xi = \bar{\xi}$, ce qui traduit le fait que la racine ξ est unique.

Remarque 1. Le théorème reste toujours valide si la fonction $\varphi(x)$ est définie et dérivable dans l'intervalle infini $-\infty < x < +\infty$ et si l'inégalité (6) est vérifiée pour tout x.

Remarque 2. Sous les conditions du théorème 1, la méthode des approximations successives converge *quel que soit le choix de la valeur initiale* x_0 dans $[a, b]$. Il en résulte que cette méthode est *autocorrectrice*, c'est-à-dire qu'une erreur de calcul isolée, à condition que les limites du segment $[a, b]$ ne soient pas dépassées, n'influe pas sur le résultat final du fait qu'une valeur incorrecte peut être considérée comme une nouvelle valeur initiale x_0. Il se peut que cela entraîne seulement le plus grand volume de travail. Grâce à la propriété d'autocorrection la méthode des approximations successives est une des méthodes de calcul les plus sûres. On comprend bien que les erreurs systématiques résultant de l'application de cette méthode peuvent empêcher l'obtention du résultat demandé.

Estimation de l'approximation. La formule (10) conduit à

$$|x_{n+p}-x_n| \leqslant |x_{n+p}-x_{n+p-1}|+|x_{n+p-1}-x_{n+p-2}|+ \\ +\ldots+|x_{n+1}-x_n| \leqslant q^{n+p-1}|x_1-x_0|+q^{n+p-2}|x_1-x_0|+ \\ +\ldots+q^n|x_1-x_0| = q^n|x_1-x_0|(1+q+q^2+\ldots+q^{p-1}).$$

La somme de la progression géométrique donne

$$|x_{n+p}-x_n| \leqslant q^n|x_1-x_0|\frac{1-q^p}{1-q} < \frac{q^n}{1-q}|x_1-x_0|.$$

En faisant tendre le nombre p vers l'infini et en retenant que $\lim\limits_{p\to\infty} x_{n+p} = \xi$, on a finalement:

$$|\xi-x_n| \leqslant \frac{q^n}{1-q}|x_1-x_0|. \qquad (15)$$

Il est donc clair que la convergence du processus itératif sera d'autant plus rapide que le nombre q est plus petit.

Pour évaluer une approximation il existe également une autre formule qui peut être utile dans certains cas. Soit

$$f(x) = x-\varphi(x).$$

Il est évident que $f'(x) = 1-\varphi'(x) \geqslant 1-q$. On en tire en tenant compte que $f(\xi) = 0$:

$$|x_n-\varphi(x_n)| = |f(x_n)-f(\xi)| = \\ = |x_n-\xi||f'(\bar{x}_n)| \geqslant (1-q)|x_n-\xi|,$$

où $\bar{x}_n \in (x_n, \xi)$ et, par conséquent,

$$|x_n-\xi| \leqslant \frac{|x_n-\varphi(x_n)|}{1-q}, \qquad (16)$$

d'où

$$|\xi-x_n| \leqslant \frac{|x_{n+1}-x_n|}{1-q}. \qquad (16')$$

L'utilisation de la formule (9) donne également

$$|\xi-x_n| \leqslant \frac{q}{1-q}|x_n-x_{n-1}|, \qquad (16'')$$

d'où il résulte notamment que si $q \leqslant \frac{1}{2}$, on a

$$|\xi-x_n| \leqslant |x_n-x_{n-1}|,$$

c'est-à-dire dans ce cas l'inégalité $|x_n-x_{n-1}| < \varepsilon$ entraîne donc l'inégalité

$$|\xi-x_n| < \varepsilon.$$

R e m a r q u e. D'après une opinion largement répandue, si lors de l'application de la méthode considérée deux approximations

successives x_{n-1} et x_n coïncident avec la précision imposée ε (par exemple, si pour ces approximations les m premières décimales se sont stabilisées), l'égalité $\xi \approx x_n$ se vérifie alors avec la même précision (cela veut dire, en particulier dans l'exemple considéré, que m décimales du nombre approché x_n sont exactes !). Dans le cas général, comme le montre bien la figure 32, cette affirmation est fausse. Plus même, on prouve aisément que si $\varphi'(x)$ est voisine de 1, la grandeur $|\xi - x_n|$ peut être grande, bien que la grandeur $|x_n - x_{n-1}|$ soit très petite.

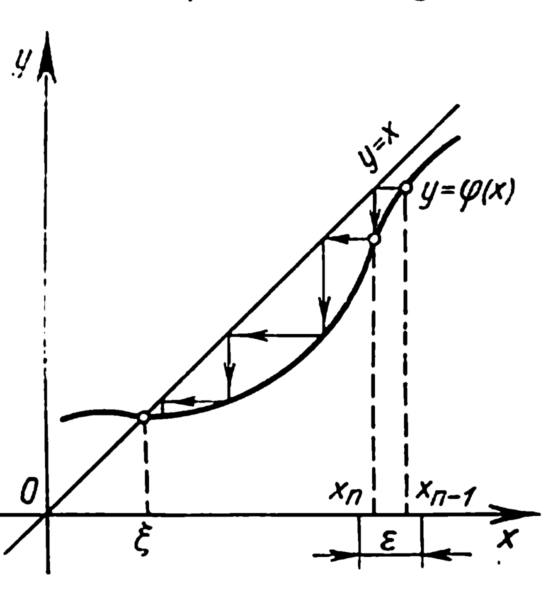


Fig. 32.

La formule (16″) permet d'évaluer l'erreur de la valeur approchée x_n d'après l'écart entre deux approximations successives x_{n-1} et x_n.

Le processus itératif doit être poursuivi tant que les deux approximations successives x_{n-1} et x_n ne vérifient l'inégalité

$$|x_n - x_{n-1}| \leqslant \frac{1-q}{q}\varepsilon,$$

où ε est la borne d'erreur absolue imposée de la racine ξ et $|\varphi'(x)| \leqslant \leqslant q$. La formule (16″) donne alors lieu à l'inégalité

$$|\xi - x_n| \leqslant \varepsilon,$$

c'est-à-dire

$$\xi = x_n \pm \varepsilon.$$

Notons que si

$$x_n = \varphi(x_{n-1})$$

et

$$\xi = \varphi(\xi),$$

on a

$$|\xi - x_n| = |\varphi(\xi) - \varphi(x_{n-1})| = |\xi - x_{n-1}|\ |\varphi'(\bar{x}_{n-1})| \leqslant \leqslant q|\xi - x_{n-1}|\ (\bar{x}_{n-1} \in (x_{n-1}, \xi)),$$

c'est-à-dire

$$|\xi - x_n| \leqslant |\xi - x_{n-1}|.$$

Ainsi, dans le cas d'un processus itératif convergent l'erreur $|\xi - x_n|$ tend monotonement vers zéro, ce qui signifie que chaque valeur suivante x_n est plus précise que la valeur antérieure x_{n-1}. Toutes ces conclusions ignorent, certes, les *e r r e u r s d'a r r o n d i*, c'est-à-dire on suppose que le calcul des approximations successives soit exact.

Dans la pratique on cherche d'abord à établir par un procédé grossier que l'équation (2) possède une racine ξ pour obtenir ensuite par la méthode itérative une approximation suffisamment précise de la valeur de cette racine, l'inégalité (6) n'étant vérifiée que dans un certain voisinage (a, b) de la racine considérée. Si le choix de la

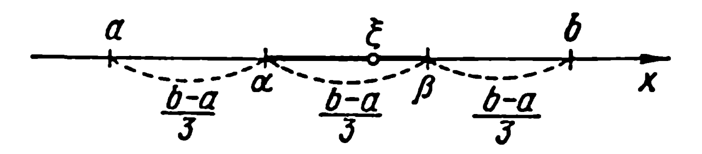


Fig. 33.

valeur initiale x_0 est mauvais, les approximations successives $x_n = \varphi(x_{n-1})$ $(n = 1, 2, \ldots)$ peuvent sortir de l'intervalle (a, b) ou même perdre tout leur sens. C'est pourquoi il convient de modifier le théorème 1.

Théorème 2. *Soit* $\varphi(x)$ *une fonction définie et dérivable sur un certain segment* $[a, b]$, *et supposons que l'équation*

$$x = \varphi(x) \tag{17}$$

ait une racine ξ *appartenant à un segment plus étroit* $[\alpha, \beta]$, *où* $\alpha = a + \frac{1}{3}(b - a)$ *et* $\beta = b - \frac{1}{3}(b - a)$ (fig. 33).

Alors, si a) $|\varphi'(x)| \leqslant q < 1$ *pour* $a < x < b$ *et si* b) *l'approximation initiale est* $x_0 \in [\alpha, \beta]$, *il vient:*

1) *toutes les approximations successives sont comprises dans l'intervalle* (a, b):

$$x_n = \varphi(x_{n-1}) \in (a, b) \quad (n = 1, 2, \ldots);$$

2) *le processus des approximations successives est convergent, c'est-à-dire il existe une limite*

$$\lim_{n\to\infty} x_n = \xi,$$

de plus, ξ *est une racine unique sur le segment* $[a, b]$ *de l'équation* (17);

3) *l'estimation* (15) *est justifiée.*

Démonstration. 1) En effet, soit

$$x_0 \in [\alpha, \beta].$$

Alors il est clair que l'égalité

$$x_1 = \varphi(x_0)$$

a un sens. En utilisant l'égalité

$$\xi = \varphi(\xi),$$

on obtient en vertu du théorème de Lagrange:

$$|x_1-\xi|=|\varphi(x_0)-\varphi(\xi)|=|x_0-\xi||\varphi'(\bar{x}_0)|\leqslant q(\beta-\alpha)<\frac{b-a}{3};$$

d'où

$$x_1 \in (a, b).$$

En général, si $x_{n-1} \in (a, b)$ $(n = 1, 2, \ldots)$ et $|x_{n-1} - \xi| < \frac{b-a}{3}$, alors

$$x_n = \varphi(x_{n-1})$$

a un sens et

$$|x_n-\xi|=|\varphi(x_{n-1})-\varphi(\xi)|=$$
$$=|x_{n-1}-\xi||\varphi'(\bar{x}_{n-1})|\leqslant q|x_{n-1}-\xi|<\frac{b-a}{3}.$$

Par conséquent, $x_n \in (a, b)$, où $n = 1, 2, 3, \ldots$

Quant aux propositions 2) et 3), leur démonstration est parfaitement analogue à celle du théorème 1.

R e m a r q u e. Supposons que dans le voisinage (a, b) de la racine ξ de l'équation (17) la dérivée $\varphi'(x)$ garde son s i g n e c o n s t a n t et que l'inégalité

$$|\varphi'(x)| \leqslant q < 1$$

soit vérifiée.

Alors, si la dérivée $\varphi'(x)$ est positive, les approximations successives

$$x_n = \varphi(x_{n-1}) \qquad (n = 1, 2, \ldots), \quad x_0 \in (a, b)$$

convergent monotonement vers la racine ξ.

Si la dérivée $\varphi'(x)$ est négative, les approximations successives oscillent autour de la racine ξ.

1) En effet, soit $0 \leqslant \varphi'(x) \leqslant q < 1$ et, par exemple,

$$x_0 < \xi.$$

Il vient

$$x_1 - \xi = \varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\,\varphi'(\xi_1) < 0,$$

où $\xi_1 \in (x_0, \xi)$; en outre,

$$|x_1 - \xi| \leqslant q|x_0 - \xi| < |x_0 - \xi|.$$

Par conséquent,

$$x_0 < x_1 < \xi.$$

En appliquant la méthode par récurrence, on a

$$x_0 < x_1 < x_2 < \ldots < \xi$$

(fig. 34a).

Si $x_0 > \xi$, on obtient un résultat analogue.

Ainsi, dans le cas d'une dérivée positive $\varphi'(x)$ il suffit de choisir l'approximation initiale x_0 dans le voisinage (a, b) de la racine ξ

intéressée; toutes les autres approximations x_n $(n = 1, 2, \ldots)$ seront comprises automatiquement dans ce voisinage et avec l'augmentation du numéro n tendront monotonement vers la racine ξ.

Fig. 34a.

2) Soit $-1 < -q \leqslant \varphi'(x) \leqslant 0$ et, par exemple, $x_0 < \xi$, de plus $x_1 = \varphi(x_0) \in (a, b)$.

On a:

$$x_1 - \xi = \varphi(x_0) - \varphi(\xi) = (x_0 - \xi)\,\varphi'(\xi_1) > 0,$$

soit $x_1 > \xi$ et $|x_1 - \xi| < |x_0 - \xi|$.

En reprenant ces raisonnements pour les approximations x_1, x_2, ..., on obtient:

$$x_0 < x_2 < \ldots < \xi < \ldots < x_3 < x_1,$$

les approximations successives étant tantôt plus petites, tantôt plus grandes que la racine ξ (fig. 34b).

Fig. 34b.

Ainsi, dans le cas d'une dérivée $\varphi'(x)$ négative, si deux approximations x_0 et x_1 appartiennent au voisinage (a, b) de la racine ξ, toute autre approximation x_n $(n = 2, 3, \ldots)$ appartient également à ce voisinage et la suite $\{x_n\}$ « e n v e l o p p e » la racine ξ.

Constatons que l'inégalité

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|$$

est évidente, ce qui traduit le fait que dans ce cas les chiffres stabilisés de l'approximation x_n appartiennent nécessairement à la racine exacte ξ.

E x e m p l e 1. Chercher les racines réelles de l'équation $x - \sin x = 0{,}25$ avec trois chiffres significatifs exacts.

S o l u t i o n. Mettons l'équation considérée sous la forme

$$x = \sin x + 0{,}25.$$

Etablissons graphiquement que sur le segment [1,1 ; 1,3] l'équation possède une racine réelle ξ égale approximativement à $x_0 = 1{,}2$ (fig. 35).

En adoptant les notations du théorème 2, posons:

$$\alpha = 1{,}1 \text{ et } \beta = 1{,}3,$$

d'où

$$a = \alpha - (\beta - \alpha) = 0{,}9 \approx \text{arc } 52°$$

et

$$b = \beta + (\beta - \alpha) = 1{,}5 \approx \text{arc } 86°.$$

Etant donné que

$$\varphi(x) = \sin x + 0{,}25$$

et

$$\varphi'(x) = \cos x,$$

pour $0{,}9 < x < 1{,}5$ on a:

$$|\varphi'(x)| \leqslant \cos 52° \approx 0{,}62 = q.$$

Si l'on choisit $x_0 \in (1{,}1\,;\ 1{,}3)$, toutes les conditions du théorème 2 seront observées et, par conséquent, il sera garanti que les approximations successives

$$x_n = \sin x_{n-1} + 0{,}25$$
$$(n = 1, 2, \ldots)$$

1) soient contenues dans l'intervalle $(0{,}9\,;\ 1{,}5)$ et 2) $x_n \to \xi$ quand $n \to \infty$.

Choisissons $x_0 = 1{,}2$ et prenons, d'après la condition du problème, la borne d'erreur absolue

$$\varepsilon = \frac{1}{2}\cdot 10^{-2}$$

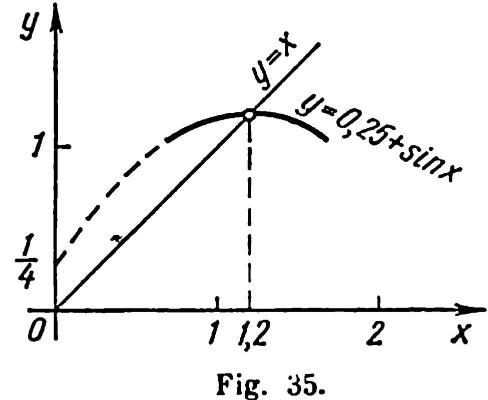


Fig. 35.

pour construire les approximations successives x_n $(n = 1, 2, \ldots)$ tant que deux approximations voisines x_{n-1} et x_n ne coïncident dans les limites de la précision égale à

$$\frac{1-q}{q}\varepsilon = 0{,}51\cdot\frac{1}{2}\cdot 10^{-2} \approx 0{,}0025.$$

On a:

$$x_1 = \sin 1{,}2 + 0{,}25 = 0{,}932 + 0{,}25 = 1{,}182;$$
$$x_2 = \sin 1{,}182 + 0{,}25 = 0{,}925 + 0{,}25 = 1{,}175;$$
$$x_3 = \sin 1{,}175 + 0{,}25 = 0{,}923 + 0{,}25 = 1{,}173;$$
$$x_4 = \sin 1{,}173 + 0{,}25 = 0{,}922 + 0{,}25 = 1{,}172;$$
$$x_5 = \sin 1{,}172 + 0{,}25 = 0{,}922 + 0{,}25 = 1{,}172.$$

La quatrième et la cinquième approximations coïncident jusqu'à quatre chiffres significatifs. Donc (cf. (16''))

$$|x_5 - \xi| \leqslant \frac{0{,}62\cdot 0{,}001}{1-0{,}62} = 0{,}0016.$$

La borne d'erreur absolue de la racine approchée x_5 (y compris l'erreur d'arrondi) ne dépassant pas

$$E = 0{,}0016 + 0{,}002 < \frac{1}{2} \cdot 10^{-2},$$

on peut poser:

$$\xi = 1{,}17 \pm 0{,}005.$$

R e m a r q u e. L'équation considérée

$$f(x) = 0 \tag{18}$$

peut être mise sous la forme de l'égalité

$$x = \varphi(x) \tag{18'}$$

en choisissant de différentes façons la fonction $\varphi(x)$.

L'écriture de (18') n'est point indifférente pour la recherche de la racine: dans certains cas $|\varphi'(x)|$ est petite dans le voisinage de ξ, dans d'autres, elle est grande. Pour la méthode des approximations successives, la représentation (18') est avantageuse si elle vérifie l'inégalité

$$|\varphi'(x)| \leqslant q < 1; \tag{19}$$

par ailleurs, plus le nombre q est petit, plus la convergence des approximations successives vers la racine ξ est rapide.

Voici un artifice suffisamment général pour ramener l'équation (18) à la forme (18') qui assure l'observation de l'inégalité (19). Supposons que la racine cherchée ξ de l'équation est comprise dans le segment $[a, b]$ et en outre

$$0 < m_1 \leqslant f'(x) \leqslant M_1 \tag{20}$$

avec $a \leqslant x \leqslant b$ *. On peut prendre notamment comme m_1 la valeur minimale de la dérivée $f'(x)$ sur le segment $[a, b]$, qui doit être positive, et comme M_1 la valeur maximale de $f'(x)$ sur le segment $[a, b]$. Remplaçons (18) par une équation équivalente

$$x = x - \lambda f(x) \quad (\lambda > 0).$$

On peut poser $\varphi(x) = x - \lambda f(x)$.

Choisissons le paramètre λ de façon que dans le voisinage considéré $[a, b]$ de la racine ξ l'inégalité

$$0 \leqslant \varphi'(x) = 1 - \lambda f'(x) \leqslant q < 1 \tag{21}$$

soit satisfaite.

L'expression (20) entraîne

$$0 \leqslant 1 - \lambda M_1 \leqslant 1 - \lambda m_1 \leqslant q.$$

* Si la dérivée $f'(x)$ est négative, au lieu de l'équation $f(x) = 0$ on considère l'équation $-f(x) = 0$.

Par conséquent, on peut adopter

$$\lambda = \frac{1}{M_1}$$

et

$$q = 1 - \frac{m_1}{M_1} < 1.$$

Ainsi l'inégalité (21) est respectée.

Exemple 2. Trouver la plus grande racine positive ξ de l'équation

$$x^3 + x = 1000 \tag{22}$$

à 10^{-4} près.

Solution. Cherchons par approximation grossière la valeur approchée de la racine $x_0 = 10$; il est clair que $\xi < x_0$.

L'équation (22) peut se mettre sous la forme

$$x = 1000 - x^3, \tag{22'}$$

ou

$$x = \frac{1000}{x^2} - \frac{1}{x}, \tag{22''}$$

ou encore

$$x = \sqrt[3]{1000 - x}, \tag{22'''}$$

etc. La plus avantageuse des variantes considérées est (22‴) parce qu'en prenant pour intervalle principal (9, 10) et en posant

$$\varphi(x) = \sqrt[3]{1000 - x},$$

on aura

$$\varphi'(x) = \frac{-1}{3\sqrt[3]{(1000 - x)^2}}.$$

D'où

$$|\varphi'(x)| \leqslant \frac{1}{3\sqrt[3]{990^2}} \approx \frac{1}{300} = q.$$

Calculons les approximations successives x_n avec un chiffre de réserve d'après les formules

$$y_n = 1000 - x_n;$$

$$x_{n+1} = \sqrt[3]{y_n} \quad (n = 0, 1, 2, \ldots).$$

Les valeurs obtenues sont portées sur le tableau 4.

Etant donné que $1 - q \approx 1$, on peut poser à 10^{-4} près $\xi = 9{,}9667$.

La méthode des approximations successives peut être appliquée également au calcul des racines des équations données sous la orme de séries entières.

Tableau 4

Valeurs des approximations successives x_n et y_n

n	x_n	y_n
0	10	990
1	9,96655	990,03345
2	9,96666	990,03334
3	9,96667	

Exemple 3. Chercher la racine réelle de l'équation [2]

$$x-\frac{x^3}{3}+\frac{x^5}{10}-\frac{x^7}{42}+\frac{x^9}{216}-\frac{x^{11}}{1320}+\ldots$$
$$\ldots+(-1)^{n-1}\frac{x^{2n-1}}{(n-1)!\,(2n-1)}+\ldots=0{,}4431135.$$

Solution. On a $x=\varphi(x)$, où

$$\varphi(x)=0{,}4431135+\frac{x^3}{3}-\frac{x^5}{10}+\frac{x^7}{42}-\frac{x^9}{216}+\frac{x^{11}}{1320}-\ldots$$

En rejetant toutes les puissances de x supérieures à la première, on trouve la valeur approchée de la racine $x_0 = 0{,}44$. Puis

$$x_1 = \varphi(0{,}44) \approx 0{,}47;$$
$$x_2 = \varphi(0{,}47) \approx 0{,}476;$$
$$x_3 = \varphi(0{,}476) \approx 0{,}4767;$$
$$x_4 = \varphi(0{,}4767) \approx 0{,}47689;$$
$$x_5 = \varphi(0{,}47689) \approx 0{,}476927;$$
$$x_6 = \varphi(0{,}476927) \approx 0{,}476934;$$
$$x_7 = \varphi(0{,}476934) \approx 0{,}476936.$$

Par conséquent, $\xi = 0{,}47693$.

Voici encore un procédé d'amélioration de la convergence du processus itératif, qui, dans certains cas, peut être utile [7].

Soit l'équation

$$x = \varphi(x)$$

telle que dans le voisinage de la racine cherchée ξ l'inégalité

$$|\varphi'(x)| \geqslant k > 1.$$

soit vraie. Pour cette équation le processus itératif est divergent. Toutefois, si on la remplace par une équation équivalente

$$x = \psi(x),$$

où $\psi(x) = \varphi^{-1}(x)$ est la fonction inverse, on obtient une équation pour laquelle le processus itératif converge du fait que

$$|\psi'(x)| = \left|\frac{1}{\varphi'(\psi(x))}\right| \leqslant \frac{1}{k} = q < 1.$$

Exemple 4. L'équation

$$f(x) \equiv x^3 - x - 1 = 0 \tag{23}$$

a une racine $\xi \in (1, 2)$, puisque $f(1) = -1 < 0$ et $f(2) = 5 > 0$.

L'équation (23) peut s'écrire

$$x = x^3 - 1. \tag{24}$$

Ici

$$\varphi(x) = x^3 - 1 \text{ et } \varphi'(x) = 3x^2,$$

c'est pourquoi

$$\varphi'(x) \geqslant 3 \text{ avec } 1 \leqslant x \leqslant 2$$

et, par conséquent, les conditions de convergence du processus itératif ne sont pas respectées.

Si l'on met l'équation (23) sous la forme

$$x = \sqrt[3]{x+1}, \tag{25}$$

on aura

$$\psi(x) = \sqrt[3]{x+1} \quad \text{et} \quad \psi'(x) = \frac{1}{3\sqrt[3]{(x+1)^2}}.$$

Il en résulte que $0 < \psi'(x) < \frac{1}{3\sqrt[3]{4}} < \frac{1}{4}$ pour $1 \leqslant x \leqslant 2$ et donc pour l'équation (25) le processus itératif converge rapidement.

§ 9. Méthode des approximations successives pour un système de deux équations

Soient deux équations à deux inconnues

$$\left.\begin{aligned} F_1(x, y) = 0, \\ F_2(x, y) = 0, \end{aligned}\right\} \tag{1}$$

dont il faut chercher les solutions réelles avec la précision demandée.

Supposons que le système (1) n'admet que des solutions isolées. Le nombre de ces solutions et leurs approximations grossières peuvent être établis en construisant les courbes $F_1(x, y) = 0$ et $F_2(x, y) = 0$ et en définissant les coordonnées de leurs points d'intersection.

Soient $x = x_0$, $y = y_0$ une solution approchée du système (1) obtenue graphiquement ou par un autre procédé quelconque (par exemple, par approximation grossière).

Voici un processus itératif qui permet dans des conditions définies d'améliorer la précision des valeurs approchées données des solutions.

A cet effet, écrivons le système (1) sous la forme

$$\left.\begin{aligned} x &= \varphi_1(x, y), \\ y &= \varphi_2(x, y) \end{aligned}\right\} \tag{2}$$

et construisons les approximations successives d'après les formules suivantes:

$$\begin{aligned} x_1 &= \varphi_1(x_0, y_0); & y_1 &= \varphi_2(x_0, y_0); \\ x_2 &= \varphi_1(x_1, y_1); & y_2 &= \varphi_2(x_1, y_1); \\ &\dots\dots\dots\dots & & \\ x_{n+1} &= \varphi_1(x_n, y_n); & y_{n+1} &= \varphi_2(x_n, y_n). \end{aligned} \tag{3}$$

Si le processus itératif (3) converge, c'est-à-dire s'il existe des limites

$$\xi = \lim_{n\to\infty} x_n \quad \text{et} \quad \eta = \lim_{n\to\infty} y_n,$$

alors, en supposant les fonctions $\varphi_1(x, y)$ et $\varphi_2(x, y)$ continues et en passant à la limite dans l'égalité (3) du type général, on obtient:

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \varphi_1(x_n, y_n),$$

$$\lim_{n\to\infty} y_{n+1} = \lim_{n\to\infty} \varphi_2(x_n, y_n).$$

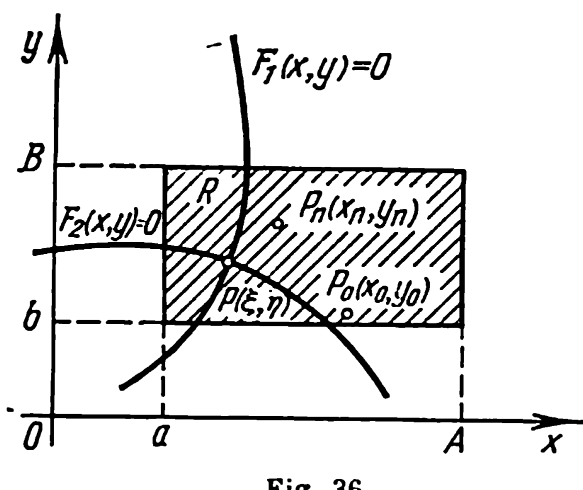


Fig. 36.

D'où

$$\xi = \varphi_1(\xi, \eta); \quad \eta = \varphi_2(\xi, \eta),$$

ce qui signifie que les valeurs limites ξ et η sont une solution du système (2) et, par conséquent, du système (1). C'est pourquoi en prenant le nombre d'itérations (3) suffisamment grand, on obtient les nombres x_n et y_n qui diffèrent aussi peu que l'on veut de la solution exacte $x = \xi$, $y = \eta$ du système (1). Le problème ainsi posé sera donc résolu. Si le processus itératif (3) est divergent, il est inutilisable.

Théorème. *Supposons que dans un certain voisinage fermé* $R\{a \leqslant x \leqslant A;\ b \leqslant y \leqslant B\}$ (fig. 36) *il existe une et seulement une solution* $x = \xi$, $y = \eta$ *du système* (2). *Si*: 1) *les fonctions* $\varphi_1(x, y)$ *et* $\varphi_2(x, y)$ *sont définies et continûment dérivables dans* R; 2) *les approximations initiales* x_0, y_0 *et toutes les approximations ultérieures*

x_n, y_n $(n = 1, 2, \ldots)$ *appartiennent à* R; 3) *les inégalités*

$$\left|\frac{\partial\varphi_1}{\partial x}\right| + \left|\frac{\partial\varphi_2}{\partial x}\right| \leqslant q_1 < 1,$$

$$\left|\frac{\partial\varphi_1}{\partial y}\right| + \left|\frac{\partial\varphi_2}{\partial y}\right| \leqslant q_2 < 1,$$

sont vérifiées dans R, *alors le processus des approximations successives* (3) *converge vers la solution* $x = \xi$, $y = \eta$ *du système* (2), *soit*

$$\lim_{n\to\infty} x_n = \xi \quad \textit{et} \quad \lim_{n\to\infty} y_n = \eta.$$

R e m a r q u e. Le théorème reste valide si la condition 3) est remplacée par 3′)

$$\left|\frac{\partial\varphi_1}{\partial x}\right| + \left|\frac{\partial\varphi_1}{\partial y}\right| \leqslant q_1 < 1,$$

$$\left|\frac{\partial\varphi_2}{\partial x}\right| + \left|\frac{\partial\varphi_2}{\partial y}\right| \leqslant q_2 < 1.$$

La démonstration de ce théorème est donnée dans [2]. Un théorème plus général est démontré dans le chapitre XIII, §§ 10 et 11.

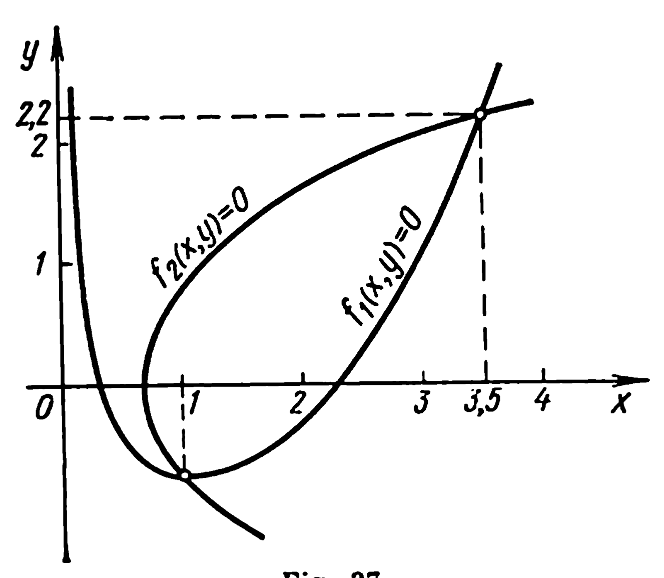


Fig. 37.

E x e m p l e. Trouver pour le système [2]

$$\left.\begin{aligned} f_1(x, y) &\equiv 2x^2 - xy - 5x + 1 = 0, \\ f_2(x, y) &\equiv x + 3\lg x - y^2 = 0 \end{aligned}\right\}$$

la solution aux coordonnées positives avec quatre chiffres significatifs exacts.

S o l u t i o n. Construisons les courbes d'équations $f_1(x, y) = 0$ et $f_2(x, y) = 0$ (fig. 37). La solution approchée qui nous inté-

resse est

$$x_0 = 3{,}5; \quad y_0 = 2{,}2.$$

Pour pouvoir appliquer la méthode des approximations successives mettons notre système sous la forme:

$$x = \sqrt{\frac{x(y+5)-1}{2}} \equiv \varphi_1(x, y);$$

$$y = \sqrt{x+3\lg x} \equiv \varphi_2(x, y).$$

Cherchons les dérivées partielles

$$\frac{\partial \varphi_1}{\partial x} = \frac{y+5}{4\sqrt{\frac{x(y+5)-1}{2}}}, \quad \frac{\partial \varphi_2}{\partial x} = \frac{1+\frac{3M}{x}}{2\sqrt{x+3\lg x}},$$

où $M = 0{,}43429$,

$$\frac{\partial \varphi_1}{\partial y} = \frac{x}{4\sqrt{\frac{x(y+5)-1}{2}}}, \quad \frac{\partial \varphi_2}{\partial y} = 0.$$

En se bornant au voisinage

$$R\{|x-3{,}5| \leqslant 0{,}1; \quad |y-2{,}2| \leqslant 0{,}1\},$$

on a:

$$\left|\frac{\partial \varphi_1}{\partial x}\right| \leqslant \frac{2{,}3+5}{4\sqrt{\frac{3{,}4(2{,}1+5)-1}{2}}} < 0{,}54;$$

$$\left|\frac{\partial \varphi_1}{\partial y}\right| \leqslant \frac{3{,}6}{4\sqrt{\frac{3{,}4(2{,}1+5)-1}{2}}} < 0{,}27;$$

$$\left|\frac{\partial \varphi_2}{\partial x}\right| \leqslant \frac{1+\frac{3\cdot 0{,}43}{3{,}4}}{2\sqrt{3{,}4+2\lg 3{,}4}} < 0{,}42;$$

$$\left|\frac{\partial \varphi_2}{\partial y}\right| = 0.$$

D'où

$$\left|\frac{\partial \varphi_1}{\partial x}\right| + \left|\frac{\partial \varphi_2}{\partial x}\right| < 0{,}54+0{,}42 = 0{,}96 < 1; \qquad (4)$$

$$\left|\frac{\partial \varphi_1}{\partial y}\right| + \left|\frac{\partial \varphi_2}{\partial y}\right| < 0{,}27+0 = 0{,}27 < 1. \qquad (5)$$

Par conséquent, si les approximations successives (x_n, y_n) ne sortent pas du domaine R (ce qu'on établit aisément au cours du calcul), le processus itératif sera convergent.

Le fait que la somme (4) est assez proche de l'unité autorise à supposer que dans le cas considéré la convergence du processus itératif sera relativement lente. Calculons les approximations successives d'après les formules

$$x_{n+1} = \sqrt{\frac{x_n (y_n + 5) - 1}{2}};$$

$$y_{n+1} = \sqrt{x_n + 3 \lg x_n} \quad (n = 0, 1, 2, \ldots).$$

Les valeurs respectives des approximations successives sont portées sur le tableau 5.

Tableau 5

Valeurs des approximations successives x_n et y_n

n	x_n	y_n
0	3,5	2,2
1	3,479	2,259
2	3,481	2,260
3	3,484	2,261
4	3,486	2,261
5	3,487	2,262
6	3,487	2,262

Ainsi, on peut poser $\xi = 3{,}487$; $\eta = 2{,}262$.

R e m a r q u e. Au lieu du processus des approximations successives (3) que nous venons d'examiner, il est parfois plus commode de faire appel au « *processus de Seidel* »:

$$x_{n+1} = \varphi_1 (x_n, y_n);$$

$$y_{n+1} = \varphi_2 (x_{n+1}, y_n) \quad (n = 0, 1, 2, \ldots).$$

La méthode des approximations successives pour des systèmes généraux fait l'objet du chapitre XIII (§§ 8 à 11).

§ 10. Méthode de Newton pour un système de deux équations

Soient x_n, y_n une solution approchée du système des équations

$$F(x, y) = 0; \quad G(x, y) = 0, \tag{1}$$

où F et G sont des fonctions continûment dérivables. En posant

$$x = x_n + h_n; \quad y = y_n + k_n,$$

on a

$$\left.\begin{aligned} F(x_n + h_n;\quad y_n + k_n) = 0, \\ G(x_n + h_n;\quad y_n + k_n) = 0. \end{aligned}\right\} \tag{2}$$

En appliquant la formule de Taylor et en se bornant aux termes linéaires par rapport à h_n et k_n, on a:

$$\left.\begin{aligned} F(x_n, y_n) + h_n F'_x(x_n, y_n) + k_n F'_y(x_n, y_n) = 0, \\ G(x_n, y_n) + h_n G'_x(x_n, y_n) + k_n G'_y(x_n, y_n) = 0. \end{aligned}\right\} \tag{3}$$

Si le jacobien

$$J(x_n, y_n) = \begin{vmatrix} F'_x(x_n, y_n) & F'_y(x_n, y_n) \\ G'_x(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix} \neq 0,$$

le système (3) amène

$$h_n = -\frac{1}{J(x_n, y_n)} \begin{vmatrix} F(x_n, y_n) & F'_y(x_n, y_n) \\ G(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix}, \tag{4}$$

$$k_n = -\frac{1}{J(x_n, y_n)} \begin{vmatrix} F'_x(x_n, y_n) & F(x_n, y_n) \\ G'_x(x_n, y_n) & G(x_n, y_n) \end{vmatrix}. \tag{5}$$

Par conséquent, on peut poser:

$$x_{n+1} = x_n - \frac{1}{J(x_n, y_n)} \begin{vmatrix} F(x_n, y_n) & F'_y(x_n, y_n) \\ G(x_n, y_n) & G'_y(x_n, y_n) \end{vmatrix}, \tag{6}$$

$$y_{n+1} = y_n - \frac{1}{J(x_n, y_n)} \begin{vmatrix} F'_x(x_n, y_n) & F(x_n, y_n) \\ G'_x(x_n, y_n) & G(x_n, y_n) \end{vmatrix} \tag{6'}$$

$$(n = 0, 1, 2, \ldots).$$

Les approximations initiales x_0, y_0 sont grossières.

Exemple. Chercher la solution réelle du système

$$\begin{aligned} F(x, y) \equiv 2x^3 - y^2 - 1 = 0; \\ G(x, y) \equiv xy^3 - y - 4 = 0. \end{aligned} \tag{1}$$

Solution. Trouvons graphiquement les approximations grossières de la solution

$$x_0 = 1{,}2; \qquad y_0 = 1{,}7.$$

En les portant dans le système (1) on obtient:

$$F(1{,}2;\ 1{,}7) = -0{,}434;$$

$$G(1{,}2;\ 1{,}7) = 0{,}1956.$$

Calculons le jacobien

$$J(x, y) = \begin{vmatrix} 6x^2 & -2y \\ y^3 & 3xy^2 - 1 \end{vmatrix};$$

d'où

$$J = \begin{vmatrix} 8{,}64 & -3{,}40 \\ 4{,}91 & 9{,}40 \end{vmatrix} = 97{,}910.$$

Calculons h_0 d'après la formule (4):

$$h_0 = -\frac{1}{97{,}910} \begin{vmatrix} -0{,}434 & -3{,}40 \\ 0{,}1956 & 9{,}40 \end{vmatrix} = \frac{3{,}389}{97{,}910} = 0{,}0349,$$

et trouvons d'après la formule (6)

$$x_1 = 1{,}2 + 0{,}0349 = 1{,}2349.$$

La formule (5) donne k_0

$$k_0 = -\frac{1}{97{,}910} \begin{vmatrix} 8{,}64 & -0{,}434 \\ 4{,}91 & 0{,}1956 \end{vmatrix} = -0{,}0390,$$

et la formule (6) permet de trouver

$$y_1 = 1{,}7 - 0{,}0390 = 1{,}6610.$$

En reprenant cette procédure avec les valeurs obtenues, on aura

$$x_2 = 1{,}2343; \quad y_2 = 1{,}6615, \text{ etc.}$$

La méthode de Newton pour les systèmes généraux est décrite dans le chapitre XIII (§§ 1 à 7).

§ 11. Application de la méthode de Newton au cas des racines complexes

Il se peut que la nécessité se présente (pour résoudre des équations différentielles linéaires, par exemple) d'améliorer la précision des racines complexes de l'équation donnée

$$f(z) = 0. \tag{1}$$

A cette fin on peut quelquefois appliquer une méthode analogue à celle de Newton.

Supposons que $f(z)$ ($z = x + iy$, $i^2 = -1$) soit une fonction analytique dans un certain voisinage U convexe * de son zéro simple isolé

$$\zeta = \xi + i\eta \qquad (f(\zeta) = 0,\ f'(\zeta) \neq 0),$$

qui est en général complexe. Soit z_n une valeur approchée de la racine qui appartient au voisinage U et

$$z_{n+1} = z_n + \Delta z_n$$

---

* C'est-à-dire deux points quelconques appartenant au voisinage U constituent les extrémités d'un segment qui appartient également à U.

une valeur exacte de la racine. En appliquant le développement en série de Taylor en z_n et en considérant que $f(z_{n+1}) \approx 0$ à Δz_n^2 près, on a

$$f(z_{n+1}) \approx f(z_n) + \Delta z_n f'(z_n) = 0;$$

d'où

$$\Delta z_n = -\frac{f(z_n)}{f'(z_n)}. \tag{2}$$

Ainsi, en partant d'une valeur quelconque z_0 on peut obtenir de proche en proche les approximations successives de la racine d'après la formule

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)} \quad (n = 0, 1, 2, \ldots). \tag{3}$$

Si $z_n \in U$ $(n = 1, 2, \ldots)$ et si la suite $\{z_n\}$ converge, la limite

$$\zeta = \lim_{n \to \infty} z_n$$

est racine de l'équation (1). En effet, en passant à la limite avec $n \to \infty$ dans l'égalité (3), on a

$$\lim_{n \to \infty} z_{n+1} = \lim_{n \to \infty} z_n - \frac{\lim_{n \to \infty} f(z_n)}{\lim_{n \to \infty} f'(z_n)}$$

ou

$$\zeta = \zeta - \frac{f(\zeta)}{f'(\zeta)}.$$

Par conséquent,

$$f(\zeta) = 0.$$

Pour évaluer l'erreur de la valeur approchée z_n supposons que

$$|f'(z)| \geqslant m_1 > 0 \quad \text{avec } z \in U.$$

Alors pour la fonction considérée

$$w = f(z)$$

il existe dans un R-voisinage suffisamment petit de la racine ξ une fonction inverse univoque

$$z = f^{-1}(w),$$

définie dans un certain voisinage $|w| < \rho$, dont on sait que sa dérivée est

$$\frac{dz}{dw} = \frac{1}{f'(z)}. \tag{4}$$

En supposant que $|f(z_n)| < \rho$, on a

$$z_n - \zeta = f^{-1}(f(z_n)) - f^{-1}(f(\zeta)) = \int_{f(\zeta)}^{f(z_n)} \frac{d}{dt}[f^{-1}(t)]\, dt = \int_{0}^{f(z_n)} \frac{dt}{f'(f^{-1}(t))}, \tag{5}$$

où t est le point variable qui parcourt le segment rectiligne entre les points $f(\zeta) = 0$ et $f(z_n)$ (fig. 38). Etant donné que $|t| < \rho$, il vient $|f^{-1}(t)| < R$ et, par conséquent,

$$|f'(f^{-1}(t))| \geqslant m_1.$$

On en tire d'après la formule (5)

$$|z_n - \zeta| \leqslant \int_0^{f(z_n)} \frac{|dt|}{|f'(f^{-1}(t))|} \leqslant \frac{|f(z_n)|}{m_1}. \quad (6)$$

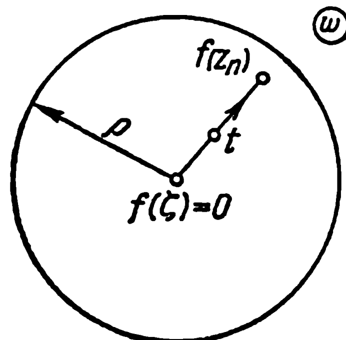


Fig. 38.

Voici sans démonstration les conditions suffisantes de l'existence d'une racine de l'équation (1) qui se déduisent du théorème d'Ostrowski [8], [9].

Théorème. *Si la fonction $f(z)$ est une fonction analytique dans un R-voisinage fermé du point z_0 et si, de plus, elle vérifie les inégalités*

1) $\left|\frac{1}{f'(z_0)}\right| \leqslant A_0$;

2) $\left|\frac{f(z_0)}{f'(z_0)}\right| \leqslant B_0 \leqslant \frac{R}{2}$;

3) $|f''(z)| \leqslant C$ avec $|z - z_0| < R$;

4) $2A_0B_0C = \mu_0 \leqslant 1$,

alors l'équation (1) *a une racine unique ζ dans le domaine $|z - z_0| \leqslant \leqslant R$ et le processus de Newton* (3) *défini par l'approximation initiale z_0 converge vers cette racine, c'est-à-dire*

$$\zeta = \lim_{n\to\infty} z_n.$$

La rapidité de la convergence du processus est caractérisée par l'estimation

$$|\zeta - z_n| \leqslant B_0 \left(\frac{1}{2}\right)^{n-1} \mu_0^{2^n-1}. \quad (7)$$

Exemple. Trouver les valeurs approchées des racines minimales en module de l'équation

$$f(z) \equiv e^z - 0{,}2z + 1 = 0. \quad (8)$$

Solution. Ici

$$f'(z) = e^z - 0{,}2.$$

Puisque $f'(z) = 0$ avec $\tilde{z} = \ln 0{,}2 \approx -1{,}79$ et

$$f(-\infty) = +\infty,\ f(\tilde{z}) > 0,\ f(+\infty) = +\infty,$$

l'équation (8) n'a pas de racine réelle.

Prenons pour approximation initiale de la racine cherchée ζ la racine z_0 au module minimal de l'équation

$$e^z + 1 = 0;$$

on peut alors poser:

$$z_0 = \pi i.$$

Déterminons les approximations successives z_n ($n = 1, 2, 3, \ldots$) de la racine ζ en appliquant la formule (3):

$$z_1 = z_0 - \frac{f(z_0)}{f'(z_0)} = \pi i - \frac{0,2\pi i}{1,2} = \frac{5}{6}\pi i = 2,618i;$$

$$z_2 = z_1 - \frac{f(z_1)}{f'(z_1)} = \frac{5\pi i}{6} - \frac{0,132 - 0,024i}{-1,868 + 0,5i} = 0,069 + 2,624i, \text{ etc.}$$

Les résultats des calculs à 0,001 près sont portés sur le tableau 6.

Tableau 6

Mise au point des racines complexes d'après la méthode de Newton

n	z_n	e^{z_n}	$f(z_n)$	$f'(z_n)$	$\Delta z_n = -\frac{f(z_n)}{f'(z_n)}$
0	$3,142i$	-1	$-0,628i$	$-1,2$	$-0,524i$
1	$2,618i$	$-0,868+0,5i$	$0,132-0,024i$	$-1,068+0,5i$	$0,153+0,040i$
2	$0,153+2,658i$	$-1,030+0,541i$	$-0,061+0,009i$	$-1,230+0,541i$	$-0,044-0,012i$
3	$0,109+2,646i$	$-0,978+0,535i$	$0+0,006i$	$-1,178+0,535i$	$-0,002+0,004i$
4	$0,107+2,650i$	$-0,981+0,525i$	$-0,002-0,005i$	$-1,181+0,525i$	$-0,000-0,004i$
5	$0,107+2,646i$	$-0,977+0,534i$	$+0,002+0,004i$	$-1,177+0,534i$	

Pour calculer e^z avec $z = x + iy$, on a fait appel à la formule connue

$$e^z = e^x (\cos y + i \sin y).$$

En posant

$$\zeta \approx z_5 = 0,107 + 2,646i,$$

on a

$$f(z_5) = 0,002 + 0,004i.$$

Si l'on considère approximativement que

$$m_1 = | f'(z_5) | \approx 1,3,$$

la formule (6) permet alors d'obtenir l'erreur

$$| \zeta - z_5 | \approx \frac{| f(z_5) |}{m_1} = \frac{0,001 \cdot \sqrt{20}}{1,3} \approx 0,004.$$

Vu que le premier membre de l'équation (8) avec des z réels prend des valeurs réelles, cette équation admet également une racine conjuguée

$$\bar{\zeta} \approx 0{,}107 - 2{,}646i,$$

égale en module à la racine ζ. En effet, on a

$$f(\bar{\zeta}) = \overline{f(\zeta)} = 0.$$

R e m a r q u e. Un autre mode de résolution de l'équation (1) consiste à la ramener à un système de deux équations réelles. En posant dans l'équation (1)

$$z = x + iy$$

et en prélevant les parties réelle et imaginaire de la fonction $f(z)$, on a:

$$f(z) \equiv u(x, y) + iv(x, y) = 0,$$

où u et v sont des fonctions réelles. On en tire que l'équation (1) est équivalente au système

$$\left.\begin{aligned} u(x, y) &= 0, \\ v(x, y) &= 0. \end{aligned}\right\} \qquad (9)$$

L'amélioration de la précision des solutions du système du type (9) est exposée aux §§ 9 et 10. Constatons que ce nouveau procédé convient également dans le cas d'une fonction $f(z)$ non analytique.

BIBLIOGRAPHIE

1. *I. Bésikovitch.* Calculs approchés. Gostekhizdat, 6e éd., 1949, chapitre VI.
2. *J. B. Scarborough.* Numerical Mathematical Analysis. John Hopkins, 1950.
3. *E. Whittaker* et *G. Robinson.* The calcul of observations. A treatise on numerical mathematics. Blacke and Son, Ltd, London and Glasgow, 1944.
4. *G. Fichtengoltz.* Cours de calcul différentiel et intégral, t. I. Gostekhizdat, 1957, chapitre IV.
5. *G. Tolstov.* Cours d'analyse mathématique, t. I. Gostekhizdat, 1954, chapitre VII.
6. *A. Guelfond.* Calcul des différences finies. Dunod, Paris, 1962, chapitre V.
7. *D. Ventsel, E. Ventsel.* Eléments de la théorie des calculs approchés. Editions de l'Académie militaire de l'Air N. Joukovski, 1949, chapitre 3, § 4.
8. *A. Ostrowski.* Recueil mathématique, 2, (1937).
9. *L. Kantorovitch.* Sur la méthode de Newton. Travaux de l'Institut mathématique V. Stéklov, XXVIII (1949), pp. 104-144.

CHAPITRE V

PROCÉDÉS SPÉCIAUX DE RÉSOLUTION APPROCHÉE DES ÉQUATIONS ALGÉBRIQUES

§ 1. Généralités

Considérons l'*équation algébrique* de degré n $(n \geqslant 1)$

$$P(x) \equiv a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0, \qquad (1)$$

où les coefficients $a_0, a_1, \ldots, a_n$ sont des nombres réels, en outre

$$a_0 \neq 0.$$

Dans le cas général la variable x est supposée complexe.

Théorème fondamental de l'algèbre. *Une équation algébrique de degré n (1) (et, par suite, un polynôme $P(x)$) admet exactement n racines réelles ou complexes, chaque racine étant prise avec son ordre de multiplicité* [1], [2].

On dit que l'ordre de multiplicité de la racine ξ de l'équation (1) est s (c'est-à-dire ξ est une *racine d'ordre de multiplicité* s), si

$$P(\xi) = P'(\xi) = \ldots = P^{(s-1)}(\xi) = 0,$$
$$P^{(s)}(\xi) \neq 0. \qquad (2)$$

Les racines complexes de l'équation (1) jouissent de la propriété d'*être conjuguées deux à deux.*

Théorème 1. *Si les coefficients d'une équation algébrique* (1) *sont réels, ses racines complexes sont conjuguées deux à deux, c'est-à-dire si $\xi = \alpha + i\beta$ (α, β étant réelles) est une racine d'ordre de multiplicité s de l'équation* (1), *le nombre $\bar{\xi} = \alpha - i\beta$ est également une racine de cette équation et son ordre de multiplicité est également s.*

Notons que les modules de ces racines sont les mêmes:

$$|\xi| = |\bar{\xi}| = \sqrt{\alpha^2 + \beta^2}.$$

Corollaire. *Une équation algébrique de degré impair à coefficients réels a au moins une racine réelle.*

Il n'est pas difficile de donner une approximation grossière aux modules des racines de l'équation (1).

Théorème 2. *Soit*

$$A = \max\{|a_1|, |a_2|, \ldots, |a_n|\},$$

où a_k *sont les coefficients de l'équation* (1).

Le module de toute racine x_k ($k = 1, \ldots, n$) *de l'équation* (1) *vérifie alors l'inégalité*

$$|x_k| < 1 + \frac{A}{|a_0|}, \qquad (3)$$

c'est-à-dire les racines de cette équation dans le plan complexe $\xi O \eta$ ($x = \xi + i\eta$) *se situent à l'intérieur du cercle*

$$|x| < 1 + \frac{A}{|a_0|} = R$$

(fig. 39).

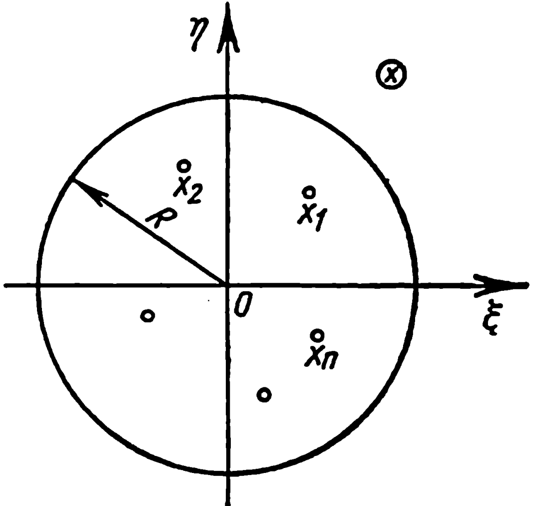


Fig. 39.

Démonstration. En posant $|x| > 1$, la formule (1) entraîne

$$|P(x)| \geqslant |a_0 x^n| - (|a_1 x^{n-1}| + |a_2 x^{n-2}| + \ldots + |a_n|) \geqslant$$
$$\geqslant |a_0|\,|x|^n - A(|x|^{n-1} + |x|^{n-2} + \ldots + 1) =$$
$$= |a_0|\,|x|^n - A\frac{|x|^n - 1}{|x| - 1} > \left(|a_0| - \frac{A}{|x|-1}\right)|x|^n.$$

Si

$$|a_0| - \frac{A}{|x|-1} \geqslant 0,$$

c'est-à-dire si

$$|x| \geqslant 1 + \frac{A}{|a_0|}, \qquad (4)$$

on en tire que

$$|P(x)| > 0.$$

Ainsi les valeurs de x qui vérifient l'inégalité (4) ne sont pas notoirement racines de l'équation (1). Par conséquent, toute racine x_k de l'équation (1) satisfait à l'inégalité opposée

$$|x_k| < 1 + \frac{A}{|a_0|}.$$

Corollaire. *Soit* $a_n \neq 0$ *et*

$$B = \max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}.$$

Toute racine x_k $(k = 1, 2, \ldots, n)$ *de l'équation* (1) *vérifie alors l'inégalité*

$$|x_k| > \frac{1}{1+\frac{B}{|a_n|}} = r, \tag{5}$$

c'est-à-dire les racines de l'équation (1) *sont comprises dans l'anneau circulaire*

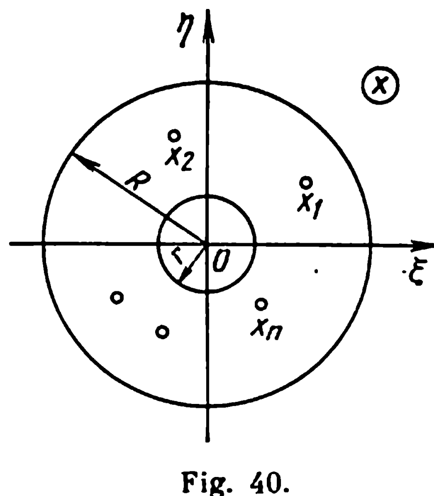


Fig. 40.

$$r < |x| < R$$

(fig. 40).

En effet, si l'on pose

$$x = \frac{1}{y},$$

on a

$$P(x) = \frac{1}{y^n} Q(y),$$

où

$$Q(y) = a_n y^n + a_{n-1} y^{n-1} + \ldots + a_0.$$

Les racines $y_k = \frac{1}{x_k}$ $(k = 1, \ldots, n)$ du polynôme $Q(y)$ vérifient, en vertu du théorème ci-dessus, l'inégalité

$$|y_k| = \frac{1}{|x_k|} < 1 + \frac{B}{|a_n|},$$

d'où

$$|x_k| > \frac{1}{1+\frac{B}{|a_n|}} = r \qquad (k = 1, \ldots, n).$$

R e m a r q u e. Les nombres r et R sont respectivement les limites *inférieure* et *supérieure* des racines positives de l'équation (1).

D'une façon analogue les nombres $-R$ et $-r$ sont les limites inférieure et supérieure des racines négatives de l'équation (1).

Si

$$x_1, x_2, \ldots, x_n$$

sont les racines de l'équation (1), son premier membre admet le développement

$$P(x) = a_0(x - x_1)(x - x_2) \ldots (x - x_n). \tag{6}$$

Après avoir multiplié entre eux les binômes de la formule (6) et égalé les coefficients des mêmes puissances de x dans les deux

membres de l'égalité (6), on en tire les relations entre les racines et les coefficients d'une équation algébrique

$$\left.\begin{aligned} &x_1+x_2+\ldots+x_n=-\frac{a_1}{a_0}, \\ &x_1x_2+x_1x_3+\ldots+x_{n-1}x_n=\frac{a_2}{a_0}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &x_1x_2\ldots x_n=(-1)^n\frac{a_n}{a_0}. \end{aligned}\right\} \qquad (7)$$

Les premiers membres des égalités (7) sont les sommes des produits des combinaisons une à une, deux à deux, etc., des racines de l'équation (1).

Exemple 1. Les racines x_1, x_2, x_3 de l'équation du troisième degré

$$x^3+px^2+qx+r=0$$

satisfont aux conditions:

$$\left.\begin{aligned} &x_1+x_2+x_3=-p, \\ &x_1x_2+x_1x_3+x_2x_3=q, \\ &x_1x_2x_3=-r. \end{aligned}\right\}$$

Si l'on tient compte de l'ordre de multiplicité des racines, le développement (6) devient

$$P(x)=a_0(x-x_1)^{\alpha_1}(x-x_2)^{\alpha_2}\ldots(x-x_m)^{\alpha_m},$$

où x_1, x_2, ..., x_m ($m\leqslant n$) sont des racines différentes de l'équation (1), et α_1, α_2, ..., α_m leurs ordres de multiplicité; en outre

$$\alpha_1+\alpha_2+\ldots+\alpha_m=n.$$

La dérivée $P'(x)$ s'exprime de la façon suivante:

$$P'(x)=a_0(x-x_1)^{\alpha_1-1}(x-x_2)^{\alpha_2-1}\ldots(x-x_m)^{\alpha_m-1}Q(x),$$

où $Q(x)$ est un polynôme tel que

$$Q(x_k)\neq 0 \quad \text{avec} \quad k=1, 2, \ldots, m.$$

C'est pourquoi le polynôme

$$R(x)=a_0(x-x_1)^{\alpha_1-1}(x-x_2)^{\alpha_2-1}\ldots(x-x_m)^{\alpha_m-1}$$

est le plus grand commun diviseur du polynôme $P(x)$ et de sa dérivée $P'(x)$. On sait que le polynôme $R(x)$ peut s'obtenir à l'aide de l'algorithme d'Euclide [1]. La composition du quotient

$$f(x)=\frac{P(x)}{R(x)}$$

permet d'obtenir le polynôme

$$f(x) = A_0 x^m + A_1 x^{m-1} + \ldots + A_m \tag{8}$$

à coefficients réels $A_0 = a_0$, A_1, ..., A_m et dont les racines x_1, x_2, ..., x_m sont différentes.

La résolution d'une équation algébrique à racines multiples se ramène donc à la résolution d'une équation algébrique de degré


inférieur à racines distinctes.

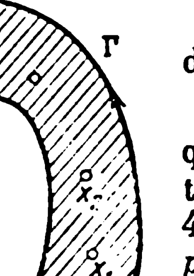


Fig. 41.

Le nombre total des racines $x_1, x_2, \ldots, x_N$ de l'équation

$$P(x) = 0,$$

qui reposent dans le plan complexe à l'intérieur d'un contour fermé simple Γ (fig. 41), peut être déterminé en partant du *principe de l'argument* [4] dont voici le sens: *si le polynôme $P(x)$ n'a pas de racines sur un contour fermé Γ, le nombre de racines N de ce polynôme à l'intérieur du contour Γ est strictement égal à l'accroissement de l'*Arg $P(x)$ *lors du parcours dans le sens positif du contour Γ, divisé par* 2π, soit

$$N = \frac{1}{2\pi} \Delta_\Gamma \operatorname{Arg} P(x),$$

chaque racine étant prise avec son ordre de multiplicité.

Si l'équation du contour Γ s'écrit

$$x = \xi(t) + i\eta(t) \qquad (0 \leqslant t \leqslant T)$$

(t étant un paramètre), pour déterminer le nombre N dans le plan XOY, on construit la courbe

$$X = X(t),\ Y = Y(t) \qquad (0 \leqslant t \leqslant T), \tag{K}$$

où

$$P(x) = P(\xi(t) + i\eta(t)) = X(t) + iY(t)$$

($X(t)$, $Y(t)$ sont des fonctions réelles), pour calculer ensuite le nombre N de tours que fait la courbe K autour de l'origine des coordonnées.

Exemple 2. Déterminer le nombre de racines de l'équation

$$P(x) \equiv x^3 - 3x + 1 = 0, \tag{9}$$

comprises à l'intérieur du cercle $|x| < 2$.

Solution. Posant

$$x = 2(\cos t + i \sin t),$$

Tableau 7

t	0	$\pm\frac{\pi}{6}$	$\pm\frac{\pi}{3}$	$\pm\frac{\pi}{2}$	$\pm\frac{2\pi}{3}$	$\pm\frac{5\pi}{6}$	$\pm\pi$
X	3	−4,22	−10	1	15	6,22	−1
Y	0	±5	±5,22	∓14	∓5,22	±5	0

on a

$$P(x) = 8(\cos t + i \sin t)^3 - 6(\cos t + i \sin t) + 1 =$$
$$= (8 \cos 3t - 6 \cos t + 1) + i(8 \sin 3t - 6 \sin t).$$

D'où

$$\left.\begin{aligned} X &= 8 \cos 3t - 6 \cos t + 1, \\ Y &= 8 \sin 3t - 6 \sin t. \end{aligned}\right\} \qquad (K)$$

Après avoir construit suivant les points la courbe K (cf. tableau 7), on voit sans peine que la courbe enveloppe trois fois l'origine des coordonnée (fig. 42). C'est pourquoi $N = 3$ et, par conséquent, l'équation (9) possède à l'intérieur du cercle $|x| < 2$ trois racines.

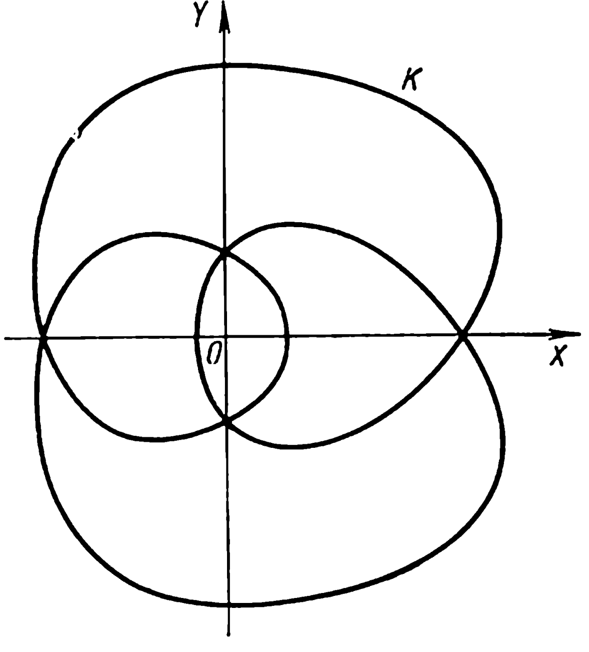


Fig. 42.

§ 2. Limites des racines réelles des équations algébriques

Dans ce paragraphe nous allons examiner les polynômes du type

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n \qquad (1)$$

à coefficients réels $a_0, a_1, \ldots, a_n$, où $a_0 \neq 0$. Nous nous proposons d'établir les limites étroites au possible des racines positives et négatives $x_1, x_2, \ldots, x_m$ $(1 \leqslant m \leqslant n)$ de l'équation

$$P(x) = 0 \qquad (2)$$

sans aborder la question de l'existence de ces racines. Notons que nous pouvons nous borner à la recherche de la limite supérieure R

seulement pour les racines positives des équations du type (2). En effet, considérons simultanément avec l'équation (2) les équations algébriques auxiliaires

$$P_1(x) \equiv x^n P\left(\frac{1}{x}\right) = 0,$$
$$P_2(x) \equiv P(-x) = 0,$$
$$P_3(x) \equiv x^n P\left(-\frac{1}{x}\right) = 0$$

et supposons que les limites supérieures de leurs racines positives sont respectivement R_1, R_2 et R_3. Alors le nombre $\frac{1}{R_1}$ constitue évidemment la limite inférieure des racines positives de l'équation (2), c'est-à-dire toute racine positive x^+ de cette équation, si elle existe, vérifie l'inégalité

$$\frac{1}{R_1} \leqslant x^+ \leqslant R.$$

D'une façon analogue les nombres $-R_2$ et $-\frac{1}{R_3}$ sont respectivement les limites inférieure et supérieure des racines négatives de l'équation (2), c'est-à-dire toute racine négative x^- de cette équation, si elle existe, vérifie l'inégalité

$$-R_2 \leqslant x^- \leqslant -\frac{1}{R_3}.$$

Indiquons certains procédés simples pour la recherche d'une limite supérieure R des racines positives de l'équation (2), en donnant certains d'entre eux sans démonstration.

Théorème de Lagrange. *Soit $a_0 > 0$ et a_k $(k \geqslant 1)$ le premier des coefficients négatifs * du polynôme $P(x)$. On peut alors prendre comme limite supérieure des racines positives de l'équation* (2) *le nombre*

$$R = 1 + \sqrt[k]{\frac{B}{a_0}}, \tag{3}$$

où B est la plus grande des valeurs absolues des coefficients négatifs du polynôme $P(x)$.

Démonstration. Soit $x > 1$. Si tout coefficient non négatif $a_1, \ldots, a_{k-1}$ du polynôme $P(x)$ est remplacé par un zéro, et tout autre coefficient $a_k, a_{k+1}, \ldots, a_n$ par un nombre négatif $-B$, la valeur du polynôme (1) ne peut que diminuer et donner lieu

* Si le coefficient de ce type n'existe pas, c'est-à-dire si tout coefficient du polynôme $P(x)$ est non négatif, le polynôme $P(x)$ n'a pas de racines positives.

à l'inégalité

$$P(x) \geqslant a_0 x^n - B(x^{n-k} + x^{n-k-1} + \ldots + 1) = a_0 x^n - B\frac{x^{n-k+1}-1}{x-1}.$$

Il en résulte avec $x > 1$

$$P(x) > a_0 x^n - \frac{B}{x-1} x^{n-k+1} = \frac{x^{n-k+1}}{x-1}[a_0 x^{k-1}(x-1) - B] >$$

$$> \frac{x^{n-k+1}}{x-1}[a_0 (x-1)^k - B].$$

Par conséquent, pour

$$x \geqslant 1 + \sqrt[k]{\frac{B}{a_0}} = R$$

on a

$$P(x) > 0,$$

ce qui signifie que toute racine positive x^+ de l'équation (2) vérifie l'inégalité

$$x^+ < R.$$

§ 3. Méthode des sommes alternées

L'idée de la méthode de Lagrange peut être généralisée de la façon suivante: soit le polynôme $P(x)$ rangé d'après les puissances décroissantes de la variable x, son coefficient du terme principal $a_0 > 0$. Mettons $P(x)$ sous forme d'une somme alternée

$$P(x) = Q_1(x) - Q_2(x) + Q_3(x) - Q_4(x) + \ldots$$
$$\ldots + Q_{2m-1}(x) - Q_{2m}(x),$$

où $Q_1(x)$ est la somme des termes consécutifs du polynôme $P(x)$ à coefficients positifs à partir de $a_0 x^n$, $-Q_2(x)$ la somme des termes consécutifs du polynôme $P(x)$ à coefficients négatifs adhérant immédiatement aux termes de la première somme, etc., le dernier terme $-Q_{2m}(x)$ étant composé d'éléments à coefficients négatifs ou étant identiquement nul.

Désignons par c_j ($j = 1, 2, \ldots, m$) les nombres positifs tels que

$$Q_{2j-1}(c_j) - Q_{2j}(c_j) \geqslant 0 \qquad (1)$$

($j = 1, 2, \ldots, m$). On peut alors admettre comme limite supérieure des racines positives de l'équation (2) (cf. § 2) le nombre

$$R = \max(c_1, c_2, \ldots, c_m). \qquad (2)$$

En effet, posons:

$$Q_{2j-1}(x) - Q_{2j}(x) = b_1^{(j)} x^{n_j} + b_2^{(j)} x^{n_j-1} + \ldots + b_p^{(j)} x^{n_j-p+1} -$$
$$- b_{p+1}^{(j)} x^{n_j-p} - b_{p+2}^{(j)} x^{n_j-p-1} - \ldots - b_{p+q}^{(j)} x^{n_j-p-q+1},$$

où

$$b_s^{(j)} \geqslant 0 \qquad (s=1, 2, \ldots, p+q),$$

en outre, $b_1^{(j)} > 0$ $(j=1, 2, \ldots, m)$.

En posant $x > 0$, on a

$$Q_{2j-1}(x) - Q_{2j}(x) = x^{n_j-p+1}\Big[(b_1^{(j)}x^{p-1} + b_2^{(j)}x^{p-2} + \ldots + b_p^{(j)}) - \\ - \Big(\frac{b_{p+1}^{(j)}}{x} + \frac{b_{p+2}^{(j)}}{x^2} + \ldots + \frac{b_{p+q}^{(j)}}{x^q}\Big)\Big]. \quad (3)$$

Il vient de la formule (3) que les fonctions $Q_{2j-1}(x) - Q_{2j}(x)$ $(j=1, 2, \ldots, m)$ croissent avec l'augmentation de x. Par conséquent, pour $x > c_j > 0$, on a

$$Q_{2j-1}(x) - Q_{2j}(x) > Q_{2j-1}(c_j) - Q_{2j}(c_j) \geqslant 0.$$

On en tire pour $x > R$

$$P(x) = \sum_{j=1}^{m} [Q_{2j-1}(x) - Q_{2j}(x)] > 0,$$

donc toutes les racines positives x^+ de l'équation (2) du § 2 vérifient la condition

$$x^+ \leqslant R.$$

Exemple. Déterminer les limites des racines réelles de l'équation

$$2x^5 - 100x^2 + 2x - 1 = 0. \quad (4)$$

Solution. Ici $a_0 = 2$ et $A = \max(100, 2, 1) = 100$. D'après le théorème 2 du § 1 la limite supérieure R des racines positives de l'équation (4) s'écrit donc

$$R = 1 + \frac{A}{a_0} = 1 + \frac{100}{2} = 51.$$

En appliquant le théorème de Lagrange et en tenant compte du fait que

$$a_k = a_3 = -100 \quad \text{et} \quad B = \max(100, 1) = 100,$$

on obtient pour la limite supérieure des racines positives une estimation bien meilleure

$$R = 1 + \sqrt[3]{\frac{100}{2}} = 1 + \sqrt[3]{50} \approx 4{,}7.$$

Enfin, en utilisant la méthode des sommes alternées, on trouve

$$2x^5 - 100x^2 = 2x^2(x^3 - 50) > 0$$

pour $x > \sqrt[3]{50}$ (par exemple pour $x > 3,7$) et

$$2x - 1 = 2\left(x - \frac{1}{2}\right) > 0 \quad \text{avec } x > 0,5.$$

Par conséquent, on peut adopter

$$R = \max(3,7;\ 0,5) = 3,7.$$

Pour déterminer la limite inférieure r des racines positives de l'équation (4), posons

$$x = \frac{1}{y}.$$

L'équation (4) se met alors sous la forme

$$y^5 - 2y^4 + 100y^3 - 2 = 0.$$

On a successivement:

$$y^5 - 2y^4 = y^4(y - 2) > 0 \quad \text{pour} \quad y > 2$$

et

$$100y^3 - 2 = 100(y^3 - 0,02) > 0 \quad \text{pour} \quad y > 0,3.$$

Par conséquent,

$$R_1 = \max(2;\ 0,3) = 2$$

et

$$r = \frac{1}{R_1} = 0,5.$$

Pour trouver la limite des racines négatives de l'équation (4) posons:

$$x = -z.$$

D'où

$$2z^5 + 10z^2 + 2z + 1 = 0. \tag{4'}$$

Les coefficients de l'équation (4') étant positifs ou nuls, cette équation n'a pas de racines positives et, par conséquent, l'équation (4) n'a pas de racines négatives.

§ 4. Méthode de Newton

Théorème de Newton. *Si pour $x = c > 0$ le polynôme $P(x)$ et toutes ses dérivées $P'(x)$, $P''(x)$, . . ., $P^{(n)}(x)$ sont non négatifs:*

$$P^{(k)}(c) \geqslant 0 \quad (k = 0, 1, 2, \ldots, n), \tag{1}$$

et $P^{(n)}(c) = n!a_0 > 0$, alors $R = c$ peut être considéré comme la limite supérieure des racines positives de l'équation

$$P(x) = 0\ . \tag{2}$$

Démonstration. Si $x > c$ et que l'on tienne compte de l'inégalité (1), la formule de Taylor entraîne

$$P(x) = P(c) + P'(c)(x-c) + \ldots + \frac{P^{(n)}(c)}{n!}(x-c)^n > 0.$$

Par conséquent, toute racine positive x^+ de l'équation (2) vérifie l'inégalité

$$x^+ \leqslant c.$$

Remarque. Pour appliquer en pratique le théorème de Newton, on cherche par la méthode des tests (en utilisant, par exemple, le schéma de Hörner) une suite croissante des nombres positifs

$$0 < c_1 \leqslant c_2 \leqslant \ldots \leqslant c_{n-1} \leqslant c_n,$$

qui vérifient les inégalités

$$P^{(n-1)}(c_1) \geqslant 0,$$
$$P^{(n-2)}(c_2) \geqslant 0,$$
$$\ldots\ldots\ldots$$
$$P'(c_{n-1}) \geqslant 0,$$
$$P(c_n) \geqslant 0.$$

Les nombres de ce type existent car on a pour $a_0 > 0$:

$$P^{(m)}(x) \to +\infty \qquad (m = 0, 1, 2, \ldots, n-1)$$

quand $x \to +\infty$. Finalement on peut admettre $c = c_n$.

En effet, puisque

$$P^{(n)}(x) = n!\,a_0 > 0,$$

la fonction $P^{(n-1)}(x)$ est croissante et, par conséquent, pour $x > c_1$ on a:

$$P^{(n-1)}(x) > P^{(n-1)}(c_1) \geqslant 0.$$

Cette dernière inégalité entraîne que la fonction $P^{(n-2)}(x)$ est croissante dans l'intervalle $[c_1, +\infty)$, donc on obtient pour $x > c_2 \geqslant c_1$:

$$P^{(n-2)}(x) > P^{(n-2)}(c_2) \geqslant 0.$$

En reprenant ce raisonnement plusieurs fois de suite, on voit enfin que $P(x)$ est une fonction croissante dans l'intervalle $[c_{n-1}, +\infty)$ et donc on a avec $x > c_n \geqslant c_{n-1}$:

$$P(x) > P(c_n) \geqslant 0.$$

Par suite, $x^+ \leqslant c_n$.

Exemple. Considérons l'équation donnée au § 3

$$P(x) = 2x^5 - 100x^2 + 2x - 1 = 0.$$

Ici

$$P'(x) = 10x^4 - 200x + 2,$$
$$P''(x) = 40x^3 - 200,$$
$$P'''(x) = 120x^2,$$
$$P^{IV}(x) = 240x,$$
$$P^{V}(x) = 240.$$

Il est clair que $P'''(x) > 0$, $P^{IV}(x) > 0$, $P^{V}(x) > 0$ pour $x > 0$. On a:

$$P''(x) = 40(x^3 - 5) > 0 \quad \text{avec} \quad x \geqslant 2.$$

Posons $c_1 = c_2 = c_3 = 2$. Puisque

$$P'(2) = 10 \cdot 16 - 200 \cdot 2 + 2 < 0,$$

on détermine le signe du nombre

$$P'(3) = 10 \cdot 81 - 200 \cdot 3 + 2 > 0.$$

On peut poser $c_4 = 3$. Ensuite, on a:

$$P(3) = 2 \cdot 243 - 100 \cdot 9 + 2 \cdot 3 - 1 < 0;$$

c'est pourquoi on calcule:

$$P(4) = 2 \cdot 1024 - 100 \cdot 16 + 2 \cdot 4 - 1 > 0.$$

Donc $c_5 = 4$. Ainsi la limite supérieure des racines positives de l'équation considérée est

$$R = 4.$$

L'estimation donnée par la méthode de Newton est plus précise que celle de Lagrange exposée dans ce qui précède, mais moins précise que l'estimation donnée par les sommes alternées (cf. exemple du § 3).

§ 5. Nombre de racines réelles d'un polynôme

Une fois qu'on a établi les limites des racines positives et négatives de l'équation algébrique

$$P(x) = 0, \tag{1}$$

où $P(x)$ est un polynôme donné, la question qui se pose est de savoir quel est le nombre de racines réelles de l'équation donnée dans un certain intervalle connu (a, b).

L'idée générale du nombre de racines réelles de l'équation (1) dans l'intervalle (a, b) est donnée par la courbe de la fonction $y = P(x)$ (fig. 43), où les racines x_1, x_2, x_3 sont les abscisses des points d'intersection de la courbe avec l'axe Ox.

Notons les propriétés simples d'un polynôme entier.

1) Si $P(a) P(b) < 0$, il y a dans l'intervalle (a, b) un nombre impair de racines du polynôme $P(x)$ qui tiennent compte de leur multiplicité;

2) Si $P(a) P(b) > 0$, soit les racines du polynôme $P(x)$ n'existent pas dans l'intervalle (a, b), soit leur nombre est pair.

La solution exhaustive du problème sur le nombre de racines réelles d'une équation algébrique dans l'intervalle considéré est donnée par la *méthode de Sturm* [1], [2].

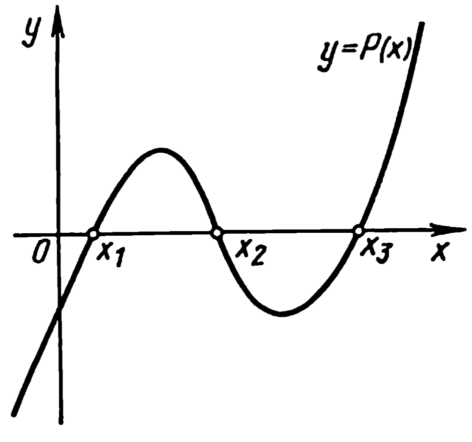


Fig. 43.

Introduisons au préalable la notion du nombre de changements de signes dans une suite numérique.

Définition. *Soit une suite finie de nombres réels différents du zéro*

$$c_1, c_2, \ldots, c_n \qquad (n \geqslant 2). \qquad (2)$$

On dit que deux éléments consécutifs c_k, c_{k+1} de la suite (2) *présentent un changement de signe si leurs signes sont contraires, c'est-à-dire si*

$$c_k c_{k+1} < 0.$$

Le nombre total de variations des signes de tous les couples d'éléments consécutifs c_k, c_{k+1} $(k = 1, 2, \ldots, n-1)$ de la suite (2), s'appelle *nombre de changements de signes* dans la suite (2).

Composons pour le polynôme considéré $P(x)$ la *suite de Sturm*

$$P(x), P_1(x), \ldots, P_2(x), \ldots, P_m(x), \qquad (3)$$

où $P_1(x) = P'(x)$, $P_2(x)$ est le reste de la division du polynôme $P(x)$ par $P_1(x)$ pris avec le signe opposé, $P_3(x)$ est le reste de la division du polynôme $P_1(x)$ par $P_2(x)$ pris avec le signe opposé, etc. Les polynômes $P_k(x)$ $(k = 2, \ldots, m)$ peuvent s'obtenir à l'aide de l'algorithme d'Euclide légèrement modifié; si le polynôme $P(x)$ n'a pas de racines multiples, le dernier élément $P_m(x)$ de la suite de Sturm est un nombre réel non nul. Notons que les éléments d'une suite de Sturm peuvent se calculer à un facteur numérique positif près.

Désignons par $N(c)$ le nombre de changements de signes de la suite de Sturm pour $x = c$, les éléments nuls de cette suite étant éliminés.

Théorème de Sturm. *Si le polynôme $P(x)$ n'a pas de racines multiples et que $P(a) \neq 0$, $P(b) \neq 0$, le nombre de ses racines réelles $N(a, b)$ dans l'intervalle $a < x < b$ est égal exactement au nombre de changements de signes perdus dans la suite de Sturm du polynôme*

$P(x)$ lors du passage de $x = a$ à $x = b$, soit

$$N(a, b) = N(a) - N(b). \tag{4}$$

Corollaire 1. Si $P(0) \neq 0$, le nombre N_+ de racines positives et le nombre N_- de racines négatives du polynôme $P(x)$ sont respectivement

$$N_+ = N(0) - N(+\infty)$$

et

$$N_- = N(-\infty) - N(0).$$

Corollaire 2. Pour que toute racine du polynôme $P(x)$ de degré n qui n'a pas de racines multiples soit réelle, il faut et il suffit que

$$N(-\infty) - N(+\infty) = n.$$

Ainsi, si

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n,$$

où $a_0 > 0$, toute racine de l'équation $P(x) = 0$ est réelle si et seulement si 1) la suite de Sturm compte le nombre maximal $n + 1$ d'éléments, c'est-à-dire si $m = n$; 2) les inégalités $P_k(+\infty) > 0$ $(k = 1, 2, \ldots, n)$ sont respectées, c'est-à-dire si le coefficient du terme principal de toutes les fonctions de Sturm $P_k(x)$ est positif [1].

Exemple. Déterminer le nombre de racines positives et négatives de l'équation

$$x^4 - 4x + 1 = 0. \tag{5}$$

Solution. La suite de Sturm s'écrit

$$P(x) = x^4 - 4x + 1;$$
$$P_1(x) = x^3 - 1;$$
$$P_2(x) = 3x - 1;$$
$$P_3(x) = 1;$$

d'où l'on tire

$$N(-\infty) = 2, \quad N(0) = 2, \quad N(+\infty) = 0.$$

Par conséquent, l'équation (5) possède

$$N_+ = 2 - 0 = 2$$

racines positives et

$$N_- = 2 - 2 = 0$$

racines négatives. Donc deux racines de l'équation (5) sont complexes.

Une suite de Sturm permet de séparer les racines d'une équation algébrique en divisant l'intervalle (a, b) contenant toutes les racines réelles de l'équation en un nombre fini d'intervalles partiels (α, β) tels que

$$N(\alpha) - N(\beta) = 1.$$

§ 6. Théorème de Budan-Fourier

La construction d'une suite de Sturm imposant en général des calculs de grande taille, pour calculer le nombre de racines réelles des équations algébriques on se borne en pratique aux procédés particuliers plus simples.

Généralisons la notion du nombre de changements de signes dans une suite numérique.

Définition. *Soit une suite finie de nombres réels*

$$c_1, c_2, \ldots, c_n, \tag{1}$$

où $c_1 \neq 0$ *et* $c_n \neq 0$.

Appelons nombre inférieur $\underline{N}$ *de changements de signes de la suite* (1) *le nombre de changements de signes de sa sous-suite obtenue en supprimant les éléments nuls.*

Appelons, d'autre part, nombre supérieur $\overline{N}$ *de changements de signes de la suite* (1) *le nombre de changements de signes de la suite* (1) *transformée de façon que tout élément nul*

$$c_k = c_{k+1} = \ldots = c_{k+l-1} = 0$$

$(c_{k-1} \neq 0,\ c_{k+l} \neq 0)$ *soit remplacé par un élément* $\breve{c}_{k+i}$ $(i = 0, 1, 2, \ldots, l-1)$ *tel que*

$$\operatorname{sgn} \breve{c}_{k+i} = (-1)^{l-i} \operatorname{sgn} c_{k+l}. \tag{2}$$

Il est évident que si la suite (1) ne possède pas d'éléments nuls, le nombre N de changements de signes de cette suite coïncide par définition avec ses nombres inférieur $\underline{N}$ et supérieur $\overline{N}$ de changements de signes:

$$N = \underline{N} = \overline{N};$$

toutefois, en général $\overline{N} \geqslant \underline{N}$.

Exemple 1. Déterminer les nombres inférieur et supérieur de changements de signes de la suite

$$1,\ 0,\ 0,\ -3,\ 1.$$

Solution. En négligeant les zéros on obtient:

$$\underline{N} = 2.$$

Pour calculer $\overline{N}$ d'après la formule (2) composons le système

$$1,\ -\varepsilon,\ \varepsilon,\ -3,\ 1,$$

où $\varepsilon > 0$. On en tire

$$\overline{N} = 4.$$

Théorème de Budan-Fourier. *Soient deux nombres* a *et* b $(a < b)$ *qui ne sont pas des racines du polynôme* $P(x)$ *de degré* n.

Alors le nombre $N(a, b)$ des racines réelles de l'équation

$$P(x) = 0, \tag{3}$$

comprises entre a et b, est égal au nombre minimal ΔN de changements de signes perdus de la suite des dérivées successives

$$P(x),\ P'(x),\ \ldots,\ P^{(n-1)}(x),\ P^{(n)}(x) \tag{4}$$

lors du passage de $x = a$ à $x = b$ ou inférieur au nombre ΔN d'un nombre pair, c'est-à-dire

$$N(a, b) = \Delta N - 2k,$$

où

$$\Delta N = \underline{N}(a) - \overline{N}(b)$$

et $\underline{N}(a)$ est le nombre inférieur de changements de signes de la suite (4) avec $x = a$, $\overline{N}(b)$ le nombre supérieur de changements de signes du système avec $x = b$ $\left(k = 0, 1, \ldots, E\left(\frac{\Delta N}{2}\right)\right)$ (cf. [1]).

On suppose que chaque racine de l'équation (3) soit prise avec son ordre de multiplicité. Si les dérivées $P^{(k)}(x)$ $(k = 1, 2, \ldots, n)$ ne s'annulent pas pour $x = a$ et $x = b$, le calcul des signes devient plus simple et, notamment,

$$\Delta N = N(a) - N(b).$$

C o r o l l a i r e 1. Si $\Delta N = 0$, entre a et b l'équation (3) n'a pas de racines réelles.

C o r o l l a i r e 2. Si $\Delta N = 1$, entre a et b l'équation (3) contient exactement une racine réelle.

R e m a r q u e. Pour calculer le nombre de changements de signes perdus ΔN de la suite (4), on fait appel au schéma de Hörner en composant deux développements:

$$P(a + h) = \alpha_0 + \alpha_1 h + \alpha_2 h^2 + \ldots + \alpha_n h^n \tag{5}$$

et

$$P(b + h) = \beta_0 + \beta_1 h + \beta_2 h^2 + \ldots + \beta_n h^n. \tag{6}$$

Soit $\underline{N}(a)$ le nombre inférieur de changements de signes des coefficients du développement (5) et, respectivement, $\overline{N}(b)$, le nombre supérieur de changements de signes des coefficients du développement (6). Etant donné que

$$\alpha_k = \frac{P^{(k)}(a)}{k!}, \quad \beta_k = \frac{P^{(k)}(b)}{k!} \quad (k = 0, 1, 2, \ldots, n),$$

les signes des nombres α_k et β_k coïncident avec ceux du système (4) pour $x = a$ et $x = b$. Donc

$$\Delta N = \underline{N}(a) - \overline{N}(b).$$

Exemple 2. Déterminer le nombre de racines réelles de l'équation

$$P(x) \equiv x^3 - x^2 + 2x - 3 = 0 \qquad (7)$$

dans l'intervalle (0, 2).

Solution. Ici $N(0)$ est évidemment le nombre de changements de signes de la suite

$$-3,\ 2,\ -1,\ 1,$$

c'est-à-dire

$$N(0) = 3.$$

Le développement $P(2+h)$ s'obtient en appliquant le schéma de Hörner

$$\begin{array}{cccc|l}
1 & -1 & 2 & +3 & 2 \\
 & 2 & 2 & 8 & \\
\hline
1 & 1 & 4 & \boxed{5} & \\
 & 2 & 6 & & \\
\hline
1 & 3 & \boxed{10} & & \\
 & 2 & & & \\
\hline
1 & \boxed{5} & & & \\
\boxed{1} & & & &
\end{array}$$

Par conséquent, $N(2)$ est le nombre de changements de signes de la suite

$$5,\ 10,\ 5,\ 1,$$

ce qui donne $N(2) = 0$.

On en tire

$$\Delta N = N(0) - N(2) = 3.$$

Ainsi dans l'intervalle (0, 2) l'équation (7) possède trois ou une racine réelle.

Théorème de Descartes. *Le nombre de racines positives d'une équation algébrique*

$$P(x) \equiv a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0 \qquad (a_0 \neq 0), \qquad (8)$$

chaque racine étant prise avec son ordre de multiplicité, est égal au nombre de changements de signes de la suite des coefficients

$$a_0,\ a_1,\ a_2,\ \ldots,\ a_n \qquad (9)$$

(les coefficients nuls étant omis) *ou inférieur à ce nombre d'un nombre pair.*

Le théorème de Descartes est l'application du théorème de Budan-Fourier à l'intervalle $(0, +\infty)$. En effet, comme

$$P^{(k)}(0) = k!a_{n-k} \qquad (k = 0, 1, \ldots, n),$$

la suite (9) est, à des facteurs positifs près, l'ensemble des dérivées $P^{(k)}(0)$ $(k = 0, 1, 2, \ldots, n)$ rangées suivant leurs ordres décroissants. C'est pourquoi le nombre de changements de signes de la suite (9) est égal à $\underline{N}(0)$, les coefficients nuls n'étant pas pris en considération. D'autre part, les dérivées $P^{(k)}(+\infty)$ $(k = 0, 1, 2, \ldots, n)$ ont évidemment le même signe et, par conséquent, $\overline{N}(+\infty) = 0$. On a donc:

$$\Delta N = \underline{N}(0) - \overline{N}(+\infty) = \underline{N}(0),$$

en outre, d'après le théorème de Budan-Fourier, le nombre de racines positives de l'équation (8) est soit égal à ΔN, soit lui est inférieur d'un nombre pair.

Corollaire. Si les coefficients de l'équation (8) sont non nuls, le nombre de racines négatives de cette équation prises avec leur ordre de multiplicité, est égal au nombre de permanences des signes des coefficients du système (9) ou inférieur à ce nombre d'un nombre pair.

Si on applique le théorème de Descartes au polynôme $P(-x)$, la démonstration de cette proposition est immédiate.

Indiquons encore une condition nécessaire pour que toutes les racines du polynôme soient réelles.

Théorème de Huat. *Si l'équation*

$$a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n = 0 \tag{10}$$

possède des coefficients réels et toutes ses racines sont réelles, le carré de tout coefficient non extrême de cette équation est supérieur au produit de ses coefficients voisins, c'est-à-dire

$$a_k^2 > a_{k-1}a_{k+1} \qquad (k = 1, 2, \ldots, n-1).$$

Corollaire. S il existe k tel que

$$a_k^2 \leqslant a_{k-1}a_{k+1},$$

l'équation (10) possède au moins un couple de racines complexes.

Exemple 3. Déterminer les racines de l'équation

$$x^4 + 8x^3 - 12x^2 + 104x - 20 = 0. \tag{11}$$

Solution. Etant donné que

$$(-12)^2 < 8 \cdot 104,$$

l'équation (11) compte des racines complexes et, par conséquent, le nombre de racines réelles de cette équation n'est pas supérieur à deux. La série des coefficients de l'équation (11) donne lieu à $\Delta N = 3$ changements de signes et $\Delta P = 1$ permanence des signes. On déduit du théorème de Descartes et de son corollaire, en tenant compte de la présence des racines complexes, que l'équation (11) a une racine positive, une racine négative et un couple de racines complexes.

§ 7. Principe de la méthode de Lobatchevski-Graeffe

Considérons l'équation algébrique de degré n

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0, \tag{1}$$

où $a_0 \neq 0$. Supposons que les racines $x_1, x_2, \ldots, x_n$ de l'équation (1) soient telles que

$$|x_1| \gg |x_2| \gg |x_3| \gg \ldots \gg |x_n|, \tag{2}$$

c'est-à-dire les racines ont des modules différents, chaque racine précédente étant bien plus grande en module que la racine ultérieure *. Autrement dit, le rapport de deux racines voisines quelconques, dans l'ordre de décroissance de leurs numéros, est une grandeur petite en module

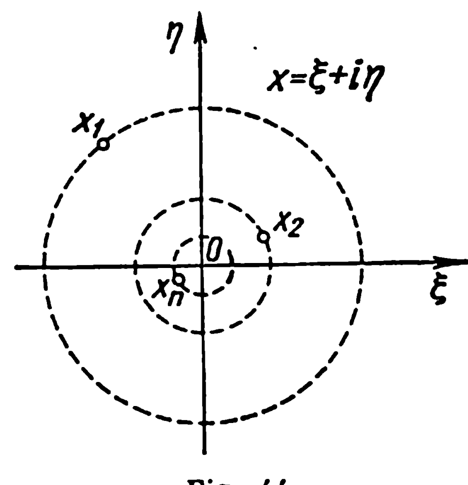


Fig. 44.

$$\left.\begin{aligned} x_2 &= \varepsilon_1 x_1, \\ x_3 &= \varepsilon_2 x_2, \\ &\ldots\ldots \\ x_n &= \varepsilon_{n-1} x_{n-1}, \end{aligned}\right\} \tag{3}$$

où $|\varepsilon_k| < \varepsilon$ et ε est une petite grandeur. Pour abréger nous dirons que les racines de ce type sont *séparées* (fig. 44).

Utilisons maintenant les relations entre les racines et les coefficients de l'équation (1) (§ 1)

$$\left.\begin{aligned} & x_1 + x_2 + \ldots + x_n = -\frac{a_1}{a_0}, \\ & x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n = \frac{a_2}{a_0}, \\ & \ldots\ldots\ldots\ldots\ldots\ldots \\ & x_1x_2 \ldots x_n = (-1)^n \frac{a_n}{a_0}. \end{aligned}\right\}$$

* Si les coefficients de l'équation (1) sont réels, la condition (2) entraîne que toutes les racines de l'équation (1) sont réelles.

Les hypothèses (3) entraînent :

$$\left.\begin{aligned} x_1(1+E_1) &= -\frac{a_1}{a_0}, \\ x_1x_2(1+E_2) &= \frac{a_2}{a_0}, \\ &\dots\dots\dots\dots\dots \\ x_1x_2\dots x_n(1+E_n) &= (-1)^n\frac{a_n}{a_0}, \end{aligned}\right\} \quad (4)$$

où $E_1, E_2, \dots, E_n$ sont des grandeurs petites en module devant l'unité. En négligeant dans les égalités (4) les grandeurs E_k $(k=1, 2, \dots, n)$, on obtient des relations approchées

$$\left.\begin{aligned} x_1 &= -\frac{a_1}{a_0}, \\ x_1x_2 &= \frac{a_2}{a_0}, \\ &\dots\dots\dots\dots \\ x_1x_2\dots x_n &= (-1)^n\frac{a_n}{a_0}. \end{aligned}\right\} \quad (5)$$

D'où les racines recherchées

$$\left.\begin{aligned} x_1 &= -\frac{a_1}{a_0}, \\ x_2 &= -\frac{a_2}{a_1} \\ &\dots\dots \\ x_n &= -\frac{a_n}{a_{n-1}}. \end{aligned}\right\} \quad (6)$$

En d'autres termes, si les racines de l'équation (1) sont séparées elles sont définies approximativement par la chaîne des équations linéaires

$$\begin{aligned} a_0x_1 + a_1 &= 0, \\ a_1x_2 + a_2 &= 0, \\ &\dots\dots \\ a_{n-1}x_n + a_n &= 0; \end{aligned}$$

par ailleurs, ces racines sont d'autant plus exactes, que dans les relations (3) les grandeurs ε_k sont plus petites en module.

Pour obtenir la séparation des racines, on compose en partant de l'équation (1) une équation transformée

$$a_0^{(m)}y^n + a_1^{(m)}y^{n-1} + \dots + a_n^{(m)} = 0, \quad (7)$$

dont les racines y_1, y_2, . . ., y_n sont les m-ièmes puissances des racines x_1, x_2, . . ., x_n de l'équation (1)

$$y_k = x_k^m \quad (k = 1, 2, \ldots, n). \tag{8}$$

Si les racines de (1) que nous considérons dans l'ordre de décroissance des modules sont différentes en module, les racines de (7) pour m suffisamment grand sont séparées du fait que

$$\frac{y_k}{y_{k-1}} = \left(\frac{x_k}{x_{k-1}}\right)^m \to 0 \text{ lorsque } m \to \infty.$$

Soit, par exemple,

$$x_1 = 2; \quad x_2 = 1,5; \quad x_3 = 1.$$

Avec $m = 100$, on a:

$$y_1 = 1,27 \cdot 10^{30}; \quad y_2 = 4,06 \cdot 10^{17}; \quad y_3 = 1$$

et donc

$$\frac{y_2}{y_1} = 3,2 \cdot 10^{-13}; \quad \frac{y_3}{y_2} = 2,5 \cdot 10^{-18}.$$

Il est d'usage de choisir pour exposant m la puissance du nombre 2, c'est-à-dire on pose $m = 2^p$, où p est un nombre naturel; la transformation elle-même se fait en p étapes dont chacune consiste à composer une équation à racines qui sont des carrés des racines de l'équation précédente.

Le calcul approché des racines y_k ($k = 1, 2, \ldots, n$) permet également de déterminer d'après les formules (8) les racines de l'équation initiale (1). La précision des calculs est d'autant plus grande que le rapport des modules des racines voisines de l'équation transformée est plus petit.

Le principe de cette méthode a été énoncé par Lobatchevski, et le schéma commode du calcul pratique est proposé par Graeffe.

Le mérite de cette méthode est que son application rend inutile la séparation des racines au sens du chapitre IV (§ 1). Il ne faut que realiser l'élimination des racines multiples par l'artifice décrit au § 1. Le calcul des racines lui-même se fait par un mode uniforme et régulier. Nous verrons plus loin que cette méthode permet également de calculer les racines complexes. Son inconvénient est la mise en œuvre de grands nombres. De plus, la vérification des calculs n'est pas assez sûre et l'estimation de la précision du résultat obtenu est plutôt difficile.

Constatons que si les racines de l'équation (1) sont distinctes, alors que les modules de certaines d'entre elles sont voisins, la convergence devient très lente. Dans ce cas il faut considérer les racines comme égales en module et recourir à des procédés de calcul spéciaux.

§ 8. Équations associées aux carrés des racines

Montrons maintenant comment composer sans peine une équation dont les racines sont les carrés des racines de l'équation algébrique donnée, pris avec le signe moins. On prend le signe moins pour rendre plus commodes les calculs en évitant au possible l'apparition des coefficients négatifs.

Pour abréger appelons *quadratisation* le processus de réduction des racines x_k $(k = 1, 2, \ldots, n)$ aux racines

$$y_k = -x_k^2. \tag{1}$$

Soit

$$P(x) \equiv a_0x^n + a_1x^{n-1} + \ldots + a_n = 0$$

l'équation donnée où $a_0 \neq 0$.

En désignant par $x_1, x_2, \ldots, x_n$ les racines de cette équation, on a:

$$P(x) = a_0(x - x_1)(x - x_2)\ldots(x - x_n).$$

Il s'ensuit

$$P(-x) = (-1)^n a_0(x + x_1)(x + x_2)\ldots(x + x_n).$$

Par conséquent

$$P(x)P(-x) = (-1)^n a_0^2(x^2 - x_1^2)(x^2 - x_2^2)\ldots(x^2 - x_n^2). \tag{2}$$

En posant

$$y = -x^2,$$

on obtient en vertu de la formule (2) le polynôme

$$Q(y) = P(x)P(-x),$$

dont les racines sont les nombres

$$y_k = -x_k^2 \qquad (k = 1, 2, \ldots, n).$$

Puisque

$$P(-x) = (-1)^n[a_0x^n - a_1x^{n-1} + a_2x^{n-2} - \ldots + (-1)^n a_n],$$

la multiplication des polynômes $P(x)$ et $P(-x)$ conduit à

$$P(x)P(-x) = (-1)^n[a_0^2x^{2n} - (a_1^2 - 2a_0a_2)x^{2n-2} + \\ + (a_2^2 - 2a_1a_3 + 2a_0a_4)x^{2n-4} - \ldots + (-1)^n a_n^2].$$

Par conséquent, l'équation qui nous intéresse s'écrit

$$Q(y) \equiv A_0y^n + A_1y^{n-1} + A_2y^{n-2} + \ldots + A_n = 0,$$

où

$$\begin{aligned} A_0 &= a_0^2, \\ A_1 &= a_1^2 - 2a_0a_2, \\ A_2 &= a_2^2 - 2a_1a_3 + 2a_0a_4, \\ &\cdots\cdots\cdots\cdots\cdots \\ A_n &= a_n^2. \end{aligned}$$

Voici une écriture plus compacte:

$$A_k = a_k^2 + 2\sum_{s=1}^{k} (-1)^s a_{k-s}a_{k+s} \quad (k = 0, 1, 2, \ldots, n),$$

où l'on suppose que $a_s = 0$ avec $s < 0$ et $s > n$.

Règle. *Chaque coefficient de l'équation transformée par la quadratisation des racines est égal au carré de l'ancien coefficient moins le double du produit des coefficients, qui lui sont voisins, plus le double du produit des coefficients adjacents à ces derniers (respectivement à gauche et à droite), etc.*, les coefficients manquants étant considérés comme nuls.

§ 9. Application de la méthode de Lobatchevski-Graeffe au cas des racines réelles distinctes

Soient les racines $x_1, x_2, \ldots, x_n$ de l'équation de degré n à coefficients réels

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \tag{1}$$

qui sont réelles et différentes en module. Rangeons-les dans l'ordre de décroissance des modules:

$$|x_1| > |x_2| > \ldots > |x_n|.$$

En appliquant successivement le processus de quadratisation des racines, composons l'équation

$$b_0y^n + b_1y^{n-1} + \ldots + b_n = 0, \tag{2}$$

dont les racines sont les nombres

$$y_k = -x_k^{2^p} \quad (k = 1, 2, \ldots, n). \tag{3}$$

Si p est suffisamment grand, les racines $y_1, y_2, \ldots, y_n$ sont séparées et d'après les résultats du § 7, elles peuvent être déterminées à partir de la chaîne des équations linéaires

$$\begin{aligned} b_0y_1 + b_1 &= 0, \\ b_1y_2 + b_2 &= 0, \\ &\cdots\cdots\cdots \\ b_{n-1}y_n + b_n &= 0. \end{aligned}$$

On en tire:

$$x_k = \pm \sqrt[2^p]{-y_k} = \sqrt[2^p]{\frac{b_k}{b_{k-1}}} \quad (k=1, 2, \ldots, n); \qquad (4)$$

les signes des racines x_k sont déterminés par une approximation grossière lors de la substitution dans l'équation considérée ou d'après les relations entre les racines et les coefficients des équations. En général, le processus de quadratisation se poursuit tant que les doubles produits ne cessent d'intervenir dans les premiers termes principaux des coefficients de l'équation transformée.

R è g l e. *Si par suite de l'annulation de doubles produits les coefficients d'une certaine équation transformée sont, dans les limites de la précision des calculs, égaux aux carrés des coefficients respectifs de l'équation transformée précédente, le processus de quadratisation des racines doit être arrêté.*

En effet, si l'équation transformée correspondant au 2^{p+1}-ième degré est de la forme

$$c_0 z^n + c_1 z^{n-1} + \ldots + c_n = 0$$

et que les relations

$$c_k = b_k^2 \quad (k = 0, 1, 2, \ldots, n),$$

soient observées, on a évidemment:

$$|x_k| = \sqrt[2^{p+1}]{\frac{c_k}{c_{k-1}}} = \sqrt[2^p]{\frac{b_k}{b_{k-1}}}.$$

Ainsi, dans ces conditions nous ne pouvons pas améliorer la précision du calcul des racines.

Puisque dans le cas de l'application de la méthode de Lobatchevski-Graeffe les coefficients des équations transformées croissent en général rapidement, il est utile de dégager leurs ordres en notant les coefficients sous forme normalisée $\alpha \cdot 10^m$, où $|\alpha| < 10$ et m est un entier. Dans les calculs imposant une précision élevée, on utilise avantageusement les logarithmes (cf. [5]).

E x e m p l e. Calculer par la méthode de Lobatchevski-Graeffe les racines de l'équation

$$x^3 - 3x + 1 = 0. \qquad (5)$$

S o l u t i o n. Les résultats des calculs avec quatre chiffres significatifs sont portés sur le tableau 8.

En s'arrêtant à la 64-ième puissance des racines, on a

$$-x_1^{64} + 3{,}445 \cdot 10^{17} = 0,$$
$$-3{,}445 \cdot 10^{17} \cdot x_2^{64} + 2{,}486 \cdot 10^{29} = 0,$$
$$-2{,}486 \cdot 10^{29} \cdot x_3^{64} + 1 = 0.$$

Tableau 8

Calcul des racines réelles par la méthode de Lobatchevski-Graeffe

Puissances	x^3	x^2	x	x^0
1	1	0	—3	1
		0	9	
		6	0	
2	1	6	9	1
		36	81	
		—18	—12	
4	1	18	69	1
		$3,24\cdot10^2$	$4,761\cdot10^3$	
		$-1,38\cdot10^2$	$-0,036\cdot10^3$	
8	1	$1,86\cdot10^2$	$4,725\cdot10^3$	1
		$3,460\cdot10^4$	$2,233\cdot10^7$	
		$-0,945\cdot10^4$	0	
16	1	$2,515\cdot10^4$	$2,233\cdot10^7$	1
		$6,325\cdot10^8$	$4,986\cdot10^{14}$	
		$-0,447\cdot10^8$	0	
32	1	$5,878\cdot10^8$	$4,986\cdot10^{14}$	1
		$3,455\cdot10^{17}$	$2\ 486\cdot10^{29}$	
		$-0,010\cdot10^{17}$	0	
64	1	$3,445\cdot10^{17}$	$2,486\cdot10^{29}$	1
		$1,187\cdot10^{35}$	$6,180\cdot10^{58}$	
		0	0	
128	1	$1,187\cdot10^{35}$	$6,180\cdot10^{58}$	1

Il en résulte

$$x_1 = \pm \sqrt[64]{3,445\cdot10^{17}},$$
$$x_2 = \pm \sqrt[64]{\frac{2,486}{3,445}\cdot10^{12}},$$
$$x_3 = \pm \sqrt[64]{\frac{1}{2,486}\cdot10^{-29}}.$$

En prenant les logarithmes:

$$\lg|x_1| = \frac{1}{64}\cdot17,53719 = 0,27402,$$
$$\lg|x_2| = \frac{1}{64}\cdot11,85831 = 0,18528,$$
$$\lg|x_3| = \frac{1}{64}\cdot(-29,39550) = \bar{1},54070,$$

et, par conséquent

$$x_1 = \pm1,879;$$
$$x_2 = \pm1,532;$$
$$x_3 = \pm0,347.$$

Pour établir les signes des racines, notons que d'après la règle de Descartes, l'équation (5) a une racine négative et deux racines positives *, de plus

$$x_1 + x_2 + x_3 = 0. \tag{6}$$

Donc la racine de module maximal doit être négative et on a finalement

$$x_1 = -1{,}879,$$
$$x_2 = 1{,}532,$$
$$x_3 = 0{,}347,$$

la relation (6) étant respectée dans les limites de la précision imposée. A titre de comparaison, donnons les valeurs des racines fournies par la formule de Cardan:

$$x_1 = 2 \cos 160° = -1{,}87938;$$
$$x_2 = 2 \cos 40° = 1{,}53208;$$
$$x_3 = 2 \cos 80° = 0{,}34730.$$

Remarquons que dans notre cas le calcul des racines est un peu simplifié, car les coefficients extrêmes de l'équation sont égaux à 1. En général, pour appliquer la méthode de Lobatchevski-Graeffe on recommande de transformer au préalable l'équation de façon à rendre le coefficient du terme principal égal à un et le terme constant à ± 1 (cf. [5]).

§ 10. Méthode de Lobatchevski-Graeffe pour le cas des racines complexes

Généralisons maintenant la notion de séparation des racines. Soit les racines $x_1, x_2, \ldots, x_n$ de l'équation

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_n = 0 \tag{1}$$

qui satisfont aux conditions

$$|x_1| \geqslant |x_2| \geqslant \ldots \geqslant |x_m| \gg |x_{m+1}| \geqslant |x_{m+2}| \geqslant \ldots \geqslant |x_n| \tag{2}$$

Autrement dit, on suppose que les racines de l'équation (1) puissent être rangées en deux *catégories* (groupes):

$$x_1, x_2, \ldots, x_m \qquad (m < n)$$

et

$$x_{m+1}, x_{m+2}, \ldots, x_n,$$

* On tient compte du fait que l'équation $P(x) \equiv x^3 - 3x + 1 = 0$ a des racines positives puisque $P(0) > 0$ et $P(1) < 0$.

de façon que les modules des racines de la première catégorie soient très grands par rapport à ceux des racines de la deuxième catégorie (cf. fig. 45 où les racines se situent dans des domaines hachurés, alors que l'intérieur de l'anneau circulaire non hachuré est privé de racines).

Ecrivons les m premières relations entre les racines et les coefficients de (1):

$$x_1+x_2+\ldots+x_m+(x_{m+1}+\ldots+x_n)=-\frac{a_1}{a_0},$$

$$x_1x_2+x_1x_3+\ldots+x_{m-1}x_m+(x_mx_{m+1}+\ldots+x_{n-1}x_n)=\frac{a_2}{a_0},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_1x_2\ldots x_m+(x_1x_2\ldots x_{m-1}x_{m+1}+\ldots+x_{n-m+1}x_{n-m+2}\ldots x_n)=$$
$$=(-1)^m\frac{a_m}{a_0}.$$

Si l'on néglige dans les dernières égalités les termes entre parenthèses relativement petits en module, on obtient les relations approchées

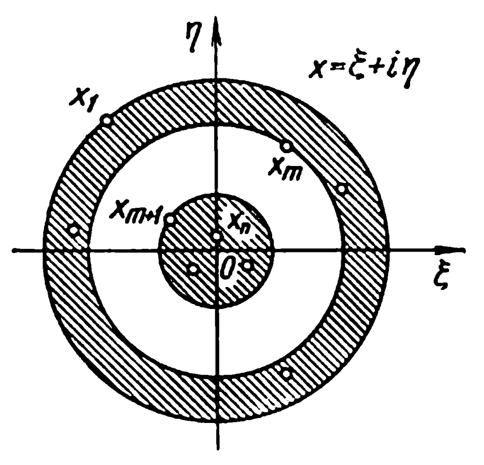


Fig. 45.

$$\left.\begin{aligned} &x_1+x_2+\ldots+x_m=-\frac{a_1}{a_0},\\ &x_1x_2+x_1x_3+\ldots+x_{m-1}x_m=\frac{a_2}{a_0},\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &x_1x_2\ldots x_m=(-1)^m\frac{a_m}{a_0}.\end{aligned}\right\}\quad(3)$$

On en déduit que les racines x_1, x_2, . . ., x_m de la première catégorie (aux grands modules) sont les racines approchées de l'équation

$$a_0x^m+a_1x^{m-1}+\ldots+a_m=0.\quad(4)$$

Les $n-m$ relations restantes entre les racines et les coefficients de l'équation (1) donnent:

$$x_1x_2\ldots x_m(x_{m+1}+x_{m+2}+\ldots+x_n)+$$
$$+x_2x_3\ldots x_{m+1}x_{m+2}+\ldots+x_{n-m}x_{n-m+1}\ldots x_n=(-1)^{m+1}\frac{a_{m+1}}{a_0},$$

$$x_1x_2\ldots x_m(x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n)+$$
$$+x_2x_3\ldots x_{m+1}x_{m+2}x_{m+3}+\ldots+x_{n-m-1}\ldots x_n=(-1)^{m+2}\frac{a_{m+2}}{a_0},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$x_1x_2\ldots x_mx_{m+1}\ldots x_n=(-1)^n\frac{a_n}{a_0}.$$

Eliminons dans les dernières égalités les termes relativement petits en module pour obtenir des relations approchées

$$x_1x_2 \ldots x_m(x_{m+1}+x_{m+2}+\ldots+x_n)=(-1)^{m+1}\frac{a_{m+1}}{a_0},$$

$$x_1x_2 \ldots x_m(x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n)=(-1)^{m+2}\frac{a_{m+2}}{a_0},$$

. . . .

$$x_1x_2 \ldots x_m x_{m+1} \ldots x_n=(-1)^n\frac{a_n}{a_0}.$$

On en tire en utilisant la dernière relation donnée par les formules (3):

$$\left.\begin{aligned} x_{m+2}+x_{m+3}+\ldots+x_n &= -\frac{a_{m+1}}{a_m},\\ x_{m+1}x_{m+2}+\ldots+x_{n-1}x_n &= \frac{a_{m+2}}{a_m},\\ &\ldots\ldots\ldots\ldots\\ x_{m+1}x_{m+2}\ldots x_n &= (-1)^{n-m}\frac{a_n}{a_m}. \end{aligned}\right\} \qquad (5)$$

Par conséquent, les racines $x_{m+1}, x_{m+2}, \ldots, x_n$ de la deuxième catégorie (aux modules petits) sont approximativement les racines de l'équation

$$a_m x^{n-m} + a_{m-1}x^{n-m-1} + \ldots + a_n = 0. \qquad (6)$$

Dans les conditions considérées, l'équation (1) se décompose ainsi en deux équations de degrés inférieurs, dont chacune définit approximativement les racines appartenant à l'une des catégories.

Un raisonnement par analogie conduit à la conclusion que si les racines de (1) peuvent former p catégories

$$x_1,\ x_2,\ \ldots,\ x_{m_1};$$

$$x_{m_1+1},\ x_{m_1+2},\ \ldots,\ x_{m_2};$$

. . . .

$$x_{m_{p-1}+1},\ x_{m_{p-1}+2},\ \ldots,\ x_{m_p}$$

$$(m_1+m_2+\ldots+m_p=n),$$

telles qu'elles vérifient la condition

$$|x_1| \geqslant |x_2| \geqslant \ldots \geqslant |x_{m_1}| \gg |x_{m_1+1}| \geqslant |x_{m_1+2}| \geqslant \ldots$$

$$\ldots \geqslant |x_{m_2}| \gg |x_{m_{p-1}+1}| \geqslant |x_{m_{p-1}+2} \geqslant \ldots \geqslant |x_{m_p}|,$$

qui, pour les modules des racines de plus bas rangs, consiste à dépasser nettement en module les racines de rangs plus élevés (ce qui nous auto-

rise de dire que ces racines sont *séparées au sens de groupe*), alors les racines de chaque catégorie peuvent être déterminées approximativement d'après les équations correspondantes

$$\left.\begin{aligned} &a_0x^{m_1}+a_1x^{m_1-1}+\ldots+a_{m_1}=0,\\ &a_{m_1}x^{m_2}+a_{m_1+1}x^{m_2-1}+\ldots+a_{m_1+m_2}=0,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &a_{m_1+m_2+\ldots+m_{p-1}}x^{m_p}+a_{m_1+m_2+\ldots+m_{p-1}+1}x^{m_p-1}+\ldots\\ &\qquad\qquad\ldots+a_{m_1+m_2+\ldots+m_p}=0, \end{aligned}\right\} \quad (7)$$

dont les degrés sont respectivement $m_1, m_2, \ldots, m_p$. En particulier, si les racines de (1) sont complètement séparées, les équations (7) sont linéaires; en l'absence des racines de même module, à un couple de racines complexes correspond dans l'ensemble des équations (7) une équation quadratique, etc.

Nous allons nous borner ici à l'examen des cas les plus simples où l'équation (1), dont les coefficients sont considérés comme réels, comporte un couple de racines complexes ou deux couples de racines complexes aux modules différents, les modules des racines réelles étant différents et autres que les modules des racines complexes. Des cas plus généraux sont décrits dans les ouvrages de Krylov [5] et de Scarborough [6].

§ 11. Cas d'un couple de racines complexes

Soit

$$\left.\begin{aligned} x_m&=u+iv,\\ x_{m+1}&=u-iv \end{aligned}\right\} \quad (1)$$

(u et v sont réels, $v \neq 0$) les racines complexes de l'équation (1) du § 10, toute autre racine x_k ($k \neq m$, $k \neq m+1$) de cette équation étant réelle et vérifiant la condition

$$|x_1| > |x_2| > \ldots > |x_m| = |x_{m+1}| > \ldots > |x_n|. \quad (2)$$

Appliquons le processus de quadratisation pour composer l'équation

$$b_0y^n + b_1y^{n-1} + \ldots + b_n = 0,$$

dont les racines sont

$$y_k = -x_k^{2^p} \qquad (k=1, 2, \ldots, n).$$

Avec un p naturel suffisamment grand, les racines réelles $y_1, \ldots$ $\ldots, y_{m-1}, y_{m+2}, \ldots, y_n$ seront séparées avec une grande précision

et pourront être déterminées à partir des équations linéaires

$$\begin{aligned} b_0 y_1 + b_1 &= 0, \\ \cdots\cdots\cdots \\ b_{m-2} y_{m-1} + b_{m-1} &= 0, \\ b_{m+1} y_{m+2} + b_{m+2} &= 0, \\ \cdots\cdots\cdots \\ b_{n-1} y_n + b_n &= 0. \end{aligned}$$

Il en résulte

$$x_k = \pm \sqrt[2^p]{\frac{b_k}{b_{k-1}}} \quad (k \neq m, \ k \neq m+1).$$

Le processus de quadratisation des racines prend fin lorsque l'étape successive fait disparaître, dans les limites de la précision imposée, les doubles produits dans les coefficients $b_1, \ldots, b_{m-1}, b_{m+1}, \ldots, b_n$. Quant au coefficient b_m, les doubles produits à disparaître en général n'en font pas partie. Il se peut même que ces produits soient supérieurs au carré et le signe du coefficient b_m soit variable. Cette circonstance est le critère de la présence des racines complexes ou des racines égales en module dans l'équation (1) du § 10, le coefficient b_m à comportement irrégulier montrant la *place* de telles racines dans la suite des modules (2).

Notons que si le coefficient b_m change de signe, l'équation considérée a des racines complexes, car dans le cas de toutes les racines réelles les coefficients des équations transformées sont évidemment non négatifs.

En vertu de la théorie générale, les racines y_m et y_{m+1}, associées aux racines complexes x_m et x_{m+1}, vérifient approximativement l'équation quadratique

$$b_{m-1} y^2 + b_m y + b_{m+1} = 0.$$

Retenons que le coefficient b_m est le coefficient *médian*. Comme

$$x_k x_{k+1} = u^2 + v^2 = r^2,$$

où

$$r = | x_k | = | x_{k+1} |$$

est le module commun des racines complexes, et

$$y_m y_{m+1} = x_k^{2^p} \cdot x_{k+1}^{2^p} = (x_k x_{k+1})^{2^p} = (r^2)^{2^p},$$

la propriété des racines d'une équation quadratique entraîne

$$(r^2)^{2^p} = \frac{b_{m+1}}{b_{m-1}} .$$

On en tire le carré du module des racines complexes

$$r^2 = \sqrt[2^p]{\frac{b_{m+1}}{b_{m-1}}}. \tag{3}$$

La partie réelle u des racines complexes s'obtient le plus simplement à partir de la relation

$$x_1 + x_2 + \ldots + x_{m-1} + (x_m + x_{m+1}) + x_{m+2} + \ldots + x_n = -\frac{a_1}{a_0}.$$

Il en résulte

$$2u = x_m + x_{m+1} = -\frac{a_1}{a_0} - \sum_{\substack{k \neq m \\ k \neq m+1}}{}' x_k$$

et, par conséquent,

$$u = -\frac{a_1}{2a_0} - \frac{1}{2} \sum_{\substack{k \neq m \\ k \neq m+1}}{}' x_k. \tag{4}$$

Le module commun r des racines complexes donné par la formule (3) permet de trouver le coefficient v de leur partie imaginaire

$$v = \sqrt{r^2 - u^2}. \tag{5}$$

Les formules (4) et (5) donnent les racines complexes cherchées

$$x_{m,\, m+1} = u \pm iv.$$

Les racines complexes peuvent être également recherchées sous une forme trigonométrique

$$x_{m,\, m+1} = r(\cos\varphi \pm i \sin\varphi).$$

E x e m p l e. Déterminer les racines de l'équation [7]

$$x^4 + x^3 - 10x^2 - 34x - 26 = 0. \tag{6}$$

S o l u t i o n. Les résultats du calcul avec quatre chiffres significatifs sont donnés par le tableau 9.

Le tableau 9 montre que les racines réelles x_1 et x_4 (dans l'ordre des modules décroissants) de la cinquième équation transformée (puissance des racines $2^5 = 32$) sont séparées. Ces racines s'obtiennent à partir des équations binômes:

$$-x_1^{32} + 2{,}005 \cdot 10^{19} = 0,$$

$$-2{,}704 \cdot 10^{43} x_4^{32} + 1{,}901 \cdot 10^{45} = 0,$$

d'où

$$x_1 = \pm \sqrt[32]{2{,}005 \cdot 10^{19}}, \quad x_4 = \pm \sqrt[32]{\frac{1{,}901}{2{,}704} \cdot 10^2}.$$

Tableau 9

Calcul des racines complexes par la méthode de Lobatchevski-Graeffe

Puissances	x^4	x^3	x^2	x	x^0
1	1	1 1 20	−10 100 68 −52	−34 1156 −520	−26
2	1	21 441 −239	116 $1,346\cdot10^4$ $-2,671\cdot10^4$ $0,135\cdot10^4$	636 $4,045\cdot10^5$ $-1,568\cdot10^5$	676
4	1	209 $4,368\cdot10^4$ $2,380\cdot10^4$	$-1,190\cdot10^4$ $1,416\cdot10^8$ $-1,035\cdot10^8$ $0,009\cdot10^8$	$2,477\cdot10^5$ $6,135\cdot10^{10}$ $1,088\cdot10^{10}$	$4,570\cdot10^5$
8	1	$6,748\cdot10^4$ $4,554\cdot10^9$ $-0,078\cdot10^9$	$3,90\cdot10^7$ $1,521\cdot10^{15}$ $-9,748\cdot10^{15}$ 0	$7,223\cdot10^{10}$ $5,216\cdot10^{21}$ $-0,016\cdot10^{21}$	$2,088\cdot10^4$
16	1	$4,476\cdot10^9$ $2,003\cdot10^{19}$ $0,002\cdot10^{19}$	$-8,227\cdot10^{15}$ $6,768\cdot10^{31}$ $-4,655\cdot10^{31}$ 0	$5,200\cdot10^{21}$ $2,704\cdot10^{43}$ 0	$4,360\cdot10^{22}$
32	1	$2,005\cdot10^{19}$ $4,020\cdot10^{38}$ 0	$2,113\cdot10^{31}$ $4,465\cdot10^{62}$ $-1,084\cdot10^{63}$ 0	$2,704\cdot10^{43}$ $7,312\cdot10^{86}$ 0	$1,901\cdot10^{45}$
64	1	$4,020\cdot10^{38}$	$-6,38\cdot10^{62}$	$7,312\cdot10^{86}$	$3,614\cdot10^{90}$

En trouvant les logarithmes, on a:

$$\lg|x_1| = \frac{1}{32}\cdot 19,30211 = 0,60319;$$

$$\lg|x_4| = \frac{1}{32}\cdot(2,27898 - 0,43201) = 0,05772.$$

Par conséquent,

$$x_1 = \pm 4,010; \quad x_4 = \pm 1,142.$$

Une approximation grossière montre que la racine x_1 est positive et la racine x_4 négative. Ainsi, on a finalement:

$$x_1 = 4,010; \quad x_4 = -1,142.$$

Le coefficient transformé affecté à x^2 changeant de sign , les racines complexes de l'équation considérée $x = x_2$ et $x = x_3$ sont définies

par l'équation trinôme

$$2{,}005\cdot 10^{19}y^2 + 2{,}113\cdot 10^{31}y + 2{,}704\cdot 10^{43} = 0,$$

où

$$y = -x^{32}.$$

D'après la théorie générale, le module des racines

$$r = |x_2| = |x_3|$$

est fourni par la formule (3)

$$r^2 = \sqrt[32]{\frac{2{,}704}{2{,}005}\cdot 10^{24}}.$$

Il vient

$$\lg r^2 = \frac{1}{32}\cdot(24{,}43201 - 0{,}30211) = 0{,}75406$$

et donc

$$r^2 = 5{,}6763.$$

En posant

$$x_2 = u + iv, \quad x_3 = u - iv,$$

la relation

$$x_1 + x_2 + x_3 + x_4 = -1$$

entraîne

$$u = \frac{1}{2}(-1 - 4{,}010 + 1{,}142) = -1{,}934.$$

Le coefficient de la partie imaginaire v est défini d'après la formule

$$v = \sqrt{r^2 - u^2} = \sqrt{5{,}6763 - 3{,}7404} = \sqrt{1{,}9359} = 1{,}395.$$

Par conséquent,

$$x_{2.3} = -1{,}934 \pm 1{,}395i.$$

Constatons que les racines x_2 et x_3 peuvent être également déterminées par les relations entre les racines et les coefficients de l'équation (6), et notamment

$$x_1 + x_2 + x_3 + x_4 = -1,$$
$$x_1x_2x_3x_4 = -26.$$

Par suite, en utilisant les valeurs x_1 et x_4 obtenues dans ce qui précède, on obtient:

$$x_2 + x_3 = -3{,}869;$$
$$x_2x_3 = 5{,}677.$$

C'est pourquoi x_2 et x_3 peuvent être trouvées comme racines d'une équation quadratique

$$x^2 + 3{,}869x + 5{,}677 = 0,$$

dont la solution donne:

$$x_{2.3} = -1{,}934 \pm 1{,}391i.$$

§ 12. Cas de deux couples de racines complexes

Soit l'équation (1) du § 10 qui admet deux couples de racines complexes:

$$x_k = u_1 + iv_1, \quad x_{k+1} = u_1 - iv_1$$

et

$$x_m = u_2 + iv_2, \quad x_{m+1} = u_2 - iv_2$$

de modules différents (u_1, v_1, u_2, v_2 réels et $v_1 \neq 0$, $v_2 \neq 0$), toutes les autres racines x_j ($j \neq k$, $j \neq k+1$, $j \neq m$, $j \neq m+1$) étant réelles, différentes entre elles en valeur absolue, non nulles * et de module différent par rapport aux racines complexes,

$$|x_1| > |x_2| > \ldots > |x_{k-1}| > |x_k| = |x_{k+1}| > \ldots > |x_m| = $$
$$= |x_{m+1}| > \ldots > |x_n| > 0. \quad (1)$$

En procédant suivant l'usage à la réduction des racines jusqu'à une certaine puissance 2^p on obtient une équation transformée

$$b_0 y^n + b_1 y^{n-1} + \ldots + b_n = 0,$$

dont les racines sont les nombres

$$y_j = -x_j^{2^p} \quad (j = 1, 2, \ldots, n).$$

Avec un p naturel suffisamment grand on découvre qu'en passant à la puissance 2^{p+1}, certains coefficients c_j de la nouvelle équation transformée

$$c_0 z^n + c_1 z^{n-1} + \ldots + c_n = 0$$

représenteront, dans les limites de la précision imposée, les carrés des coefficients correspondants b_j de l'équation transformée précédente. Sous l'hypothèse (1) on a finalement:

$$c_j = b_j^2 \text{ avec } j = 0, 1, 2, \ldots, k-1, k+1, \ldots$$
$$\ldots, m-1, m+1, \ldots, n$$

et

$$c_k \neq b_k^2 \text{ et } c_m \neq b_m^2.$$

Cette circonstance permet de déterminer la place des racines complexes. Remarquons qu'un changement de signes des coefficients b_k et b_m au cours du processus décrit pour les exposants différents 2^p est une condition suffisante de l'existence de deux couples de racines complexes de l'équation (1) du § 10.

Les racines réelles x_j de l'équation considérée se déterminent à partir des équations binômes

$$-b_{j-1} x_j^{2^p} + b_j = 0. \quad (2)$$

* Les racines nulles peuvent être mises en évidence au préalable.

Par suite

$$x_j = \pm \sqrt[2^p]{\frac{b_j}{b_{j-1}}} \quad (j \neq k,\ j \neq k+1,\ j \neq m,\ j \neq m+1).$$

Les racines complexes x_k, x_{k+1} et x_m, x_{m+1} sont déterminées respectivement à partir des équations trinômes

$$b_{k-1}x^{2^{p+1}} - b_k x^{2^p} + b_{k+1} = 0 \tag{2'}$$

et

$$b_{m-1}x^{2^{p+1}} - b_m x^{2^p} + b_{m+1} = 0. \tag{2''}$$

Introduisons les notations:

$$r_1 = |x_k| = |x_{k+1}|$$

et

$$r_2 = |x_m| = |x_{m+1}|.$$

En prenant en considération que

$$r_1^2 = x_k x_{k+1}$$

et

$$r_2^2 = x_m x_{m+1},$$

les équations (2′) et (2″) permettent de calculer les carrés des modules des racines complexes

$$r_1^2 = \sqrt[2^p]{\frac{b_{k+1}}{b_{k-1}}} \quad \text{et} \quad r_2^2 = \sqrt[2^p]{\frac{b_{m+1}}{b_{m-1}}}.$$

Pour déterminer les parties réelles u_1 et u_2 des racines complexes, on utilise les relations entre les racines et les coefficients de (1) du § 10, et notamment

$$x_2 x_3 \ldots x_n + x_1 x_3 \ldots x_n + \ldots + x_1 x_2 \ldots x_{n-1} = (-1)^{n-1} \frac{a_{n-1}}{a_0}$$

et

$$x_1 x_2 \ldots x_n = (-1)^n \frac{a_n}{a_0}.$$

Divisant la première égalité par la deuxième on obtient:

$$\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n} = -\frac{a_{n-1}}{a_n}.$$

En outre,

$$x_1 + x_2 + \ldots + x_n = -\frac{a_1}{a_0}.$$

On en tire, en tenant compte des relations

$$x_k + x_{k+1} + x_m + x_{m+1} = 2u_1 + 2u_2$$

et

$$\frac{1}{x_k}+\frac{1}{x_{k+1}}+\frac{1}{x_m}+\frac{1}{x_{m+1}}=\frac{2u_1}{r_1^2}+\frac{2u_2}{r_2^2},$$

le système linéaire d'équations suivant:

$$\left.\begin{aligned} u_1+u_2 &= -\frac{a_1}{2a_0}-\frac{1}{2}\sigma, \\ \frac{u_1}{r_1^2}+\frac{u_2}{r_2^2} &= -\frac{a_{n-1}}{2a_n}-\frac{1}{2}\sigma', \end{aligned}\right\} \tag{3}$$

où σ est la somme des racines réelles et σ' la somme de leurs grandeurs inverses:

$$\sigma=\sum_{j\neq k,\ k+1,\ m,\ m+1} x_j$$

et

$$\sigma'=\sum_{j\neq k,\ k+1,\ m,\ m+1} \frac{1}{x_j}.$$

Après avoir calculé u_1 et u_2 à partir du système (3), on détermine les coefficients v_1 et v_2 des parties imaginaires des racines d'après les formules

$$v_1=\sqrt{r_1^2-u_1^2},\quad v_2=\sqrt{r_2^2-u_2^2}.$$

Ainsi, on a finalement:

$$x_{k,\ k+1}=u_1\pm iv_1$$

et

$$x_{m,\ m+1}=u_2\pm iv_2.$$

E x e m p l e. Résoudre par la méthode de Lobatchevski-Graeffe l'équation [7]

$$x^4+4x^2-3x+3=0. \tag{4}$$

S o l u t i o n. Appliquons la méthode de quadratisation jusqu'à la puissance 16 et effectuons le calcul avec quatre chiffres significatifs exacts pour obtenir les résultats fournis par le tableau 10.

On voit sans peine que dans la transformation suivante le coefficient médian sera égal au carré de sa valeur antérieure. On arrête donc le processus. Puisque parmi les coefficients de l'équation transformée il y a, dans le cas de la puissance 16, deux coefficients négatifs, l'équation (4) admet deux couples de racines complexes:

$$x_{1,2}=u_1\pm iv_1$$

et

$$x_{3,4}=u_2\pm iv_2,$$

Tableau 10

Calcul de deux couples de racines complexes par la méthode de Lobatchevski-Graeffe

Puissances	x^4	x^3	x^2	x	x^0
1	1	0 0 −8	4 16 0 6	−3 9 −24	3
2	1	−8 64 −44	22 484 −240 18	−15 225 −396	9
4	1	20 $4\cdot10^2$ $-5{,}24\cdot10^2$	262 $6{,}864\cdot10^4$ $0{,}684\cdot10^4$ $0{,}016\cdot10^4$	−171 $2{,}924\cdot10^4$ $-4{,}244\cdot10^4$	81
8	1	$-1{,}24\cdot10^2$ $1{,}538\cdot10^4$ $-15{,}128\cdot10^4$	$7{,}564\cdot10^4$ $5{,}723\cdot10^9$ $-0{,}003\cdot10^9$ 0	$-1{,}320\cdot10^4$ $1{,}743\cdot10^8$ $-9{,}927\cdot10^8$	$6{,}561\cdot10^3$
16	1	$-1{,}359\cdot10^5$	$5{,}720\cdot10^9$	$-8{,}184\cdot10^8$	$4{,}305\cdot10^7$

qui satisfont respectivement aux équations trinômes

$$x^{32} + 1{,}359\cdot10^5\cdot x^{16} + 5{,}720\cdot10^9 = 0$$

et

$$5{,}720\cdot10^9\cdot x^{32} + 8{,}184\cdot10^8\cdot x^{16} + 4{,}305\cdot10^7 = 0.$$

On en tire les carrés des modules de ces racines:

$$r_1^2 = \sqrt[16]{5{,}720\cdot10^9} = 4{,}072$$

et

$$r_2^2 = \sqrt[16]{\frac{4{,}305}{5{,}720}\cdot10^{-2}} = 0{,}7367.$$

Puisque

$$\frac{1}{r_1^2} = 0{,}2456\,; \quad \frac{1}{r_2^2} = 1{,}3574,$$

en vertu du système (3) on obtient le système

$$u_1 + u_2 = 0,$$
$$0{,}2456u_1 + 1{,}3574u_2 = 0{,}5,$$

qui permet de rechercher les parties réelles u_1 et u_2 des racines. Il en résulte que

$$u_1 = -0{,}4497\,;$$
$$u_2 = 0{,}4497.$$

Utilisant les carrés des modules r_1^2 et r_2^2 des racines, on détermine les coefficients v_1 et v_2 des parties imaginaires des racines:

$$v_1 = \sqrt{r_1^2 - u_1^2} = 1{,}967;$$
$$v_2 = \sqrt{r_2^2 - u_2^2} = 0{,}731.$$

Donc les racines de l'équation (4) sont de la forme

$$x_{1,2} = -0{,}450 \pm 1{,}967i$$

et

$$x_{3,4} = 0{,}450 \pm 0{,}731i.$$

§ 13. Méthode de Bernoulli

Soit l'équation algébrique

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \quad (a_0 \neq 0), \tag{1}$$

dont les racines $x_1, x_2, \ldots, x_n$ sont différentes.

Utilisons les coefficients a_k $(k = 0, 1, \ldots, n)$ pour construire ce qu'on appelle *équation aux différences finies*

$$a_0y_{n+i} + a_1y_{n+i-1} + \ldots + a_ny_i = 0 \quad (i = 0, 1, 2, \ldots), \tag{2}$$

qui est une relation de récurrence associant $n + 1$ termes consécutifs quelconques d'une suite infinie

$$y_0, y_1, y_2, \ldots, y_i, \ldots \tag{3}$$

La suite (3) $y_i = f(i)$ $(i = 0, 1, 2, \ldots)$ dont les termes satisfont à l'équation aux différences finies (2) est dite *solution* de cette équation. Pour construire une solution y_i, il suffit de donner n de ses *valeurs initiales* $y_0, y_1, \ldots, y_{n-1}$; les autres termes $y_n, y_{n+1}, \ldots$ peuvent être trouvés de proche en proche d'après l'équation (2).

La théorie des différences finies [8] montre que si les racines $x_1, x_2, \ldots, x_n$ d'une équation algébrique (1) sont distinctes, toute solution de l'équation aux différences finies (2) est de la forme

$$y_i = C_1x_1^i + C_2x_2^i + \ldots + C_nx_n^i \quad (i = 0, 1, 2, \ldots), \tag{4}$$

où $C_1, C_2, \ldots, C_n$ sont des constantes arbitraires. Ainsi (1) est une équation *caractéristique* par rapport à (2). Les constantes $C_1, C_2, \ldots, C_n$ peuvent être déterminées d'après des *conditions initiales*

$$\left.\begin{aligned} y_0 &= C_1 + C_2 + \ldots + C_n, \\ y_1 &= C_1x_1 + C_2x_2 + \ldots + C_nx_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{n-1} &= C_1x_1^{n-1} + C_2x_2^{n-1} + \ldots + C_nx_n^{n-1}. \end{aligned}\right\} \tag{5}$$

Théorème. *Soit une équation algébrique* (1) *possédant une seule racine* x_1 *maximale en module. Alors le rapport de deux termes*

consécutifs y_{i+1} *et* y_i *d'une solution de l'équation aux différences finies* (2) *tend en général vers une limite égale à* x_1

$$\lim_{i\to\infty} \frac{y_{i+1}}{y_i} = x_1. \tag{6}$$

D é m o n s t r a t i o n. Soit

$$|x_1| > |x_2| \geqslant \ldots \geqslant |x_n|. \tag{7}$$

En supposant que les racines x_k ($k = 1, 2, \ldots, n$) soient différentes, la formule (4) entraîne:

$$y_i = x_1^i \left[C_1 + C_2 \left(\frac{x_2}{x_1}\right)^i + \ldots + C_n \left(\frac{x_n}{x_1}\right)^i\right]$$

et

$$y_{i+1} = x_1^{i+1} \left[C_1 + C_2 \left(\frac{x_2}{x_1}\right)^{i+1} + \ldots + C_n \left(\frac{x_n}{x_1}\right)^{i+1}\right].$$

D'où

$$\frac{y_{i+1}}{y_i} = x_1 \cdot \frac{C_1 + C_2 \left(\frac{x_2}{x_1}\right)^{i+1} + \ldots + C_n \left(\frac{x_n}{x_1}\right)^{i+1}}{C_1 + C_2 \left(\frac{x_2}{x_1}\right)^i + \ldots + C_n \left(\frac{x_n}{x_1}\right)^i}. \tag{8}$$

Si $C_1 \neq 0$, en passant à la limite dans la formule (8) pour $i \to \infty$ et en tenant compte de ce que les inégalités (7) donnent lieu aux relations limites

$$\left(\frac{x_2}{x_1}\right)^i \to 0, \ \ldots, \ \left(\frac{x_n}{x_1}\right)^i \to 0,$$

on aura:

$$\lim_{i\to\infty} \frac{y_{i+1}}{y_i} = x_1.$$

R e m a r q u e 1. Si un mauvais choix de la solution entraîne que $C_1 = 0$ et $C_2 \neq 0$, la limite (6) sera égale à la racine suivante maximale en module de l'équation (1).

R e m a r q u e 2. Si pour une solution y_i le rapport $\frac{y_{i+1}}{y_i}$ oscille sans tendre vers une limite, on peut supposer que (1) possède des racines complexes maximales en module.

R e m a r q u e 3. En effectuant dans (1) le changement de variable

$$x = \frac{1}{z},$$

on peut obtenir par la méthode de Bernoulli la racine non nulle minimale en module.

Ainsi la recherche approchée de la racine x_1 maximale en module peut se faire d'après la formule

$$x_1 \approx \frac{y_i}{y_{i-1}},$$

où i est suffisamment grand.

Pour appliquer en pratique la méthode de Bernoulli, il faut donner les nombres arbitraires $y_0, y_1, \ldots, y_{n-1}$, puis, en utilisant la formule

$$y_{n+i} = -\frac{1}{a_0}(a_n y_i + a_{n-1} y_{i-1} + \ldots + a_1 y_{n+i-1}) \quad (i = 0, 1, 2, \ldots),$$

calculer la suite des nombres $y_n, y_{n+1}, y_{n+2}, \ldots$ et les rapports $\frac{y_n}{y_{n-1}}, \frac{y_{n+1}}{y_n}, \frac{y_{n+2}}{y_{n+1}}, \ldots$ Si, pour i croissant, le rapport $\frac{y_{n+i}}{y_{n+i-1}}$ tend à s'approcher d'un certain nombre ξ, ce dernier est posé égal à la racine x_1 de (1) maximale en module. Dans le cas contraire, il se peut très bien que l'équation (1) possède plusieurs racines maximales en module ou, ce qui est moins probable, dans le système initial des nombres $y_0, y_1, \ldots, y_{n-1}$ le coefficient $C_1 = 0$.

Si l'on connaît une valeur grossière α de la racine x_1 maximale en module, il faut poser:

$$y_0 = 1, \quad y_1 = \alpha, \ \ldots, \ y_{n-1} = \alpha^{n-1},$$

afin d'accélérer la convergence.

Notons que la méthode de Bernoulli se ramène à la répétition des opérations semblables, il est donc commode de la réaliser sur des ordinateurs.

Les valeurs initiales y_i ($i = 0, 1, \ldots, n - 1$) peuvent en général être arbitraires. Dans les cas courants, on prend $y_0 = y_1 = \ldots = y_{n-2} = 0$; $y_{n-1} = 1$. F. B. Hildebrand [9] a proposé de choisir y_i de façon que tout coefficient C_i de la formule (4) soit égal à 1. Dans ce cas, s'il n'y a qu'une racine de (1) maximale en module, le processus $\frac{y_i}{y_{i-1}}$ converge quand $i \to \infty$.

La méthode de Bernoulli peut être appliquée également pour calculer les racines complexes de (1) [10].

E x e m p l e. Trouver la racine x_1 maximale en module de l'équation

$$x^5 + 5x^4 - 5 = 0.$$

S o l u t i o n. L'équation aux différences finies correspondante est de la forme

$$y_{i+5} = 5(y_i - y_{i+4}) \quad (i = 0, 1, 2, \ldots). \tag{9}$$

Prenons arbitrairement les valeurs

$$y_0 = 0, \ y_1 = 0, \ y_2 = 0, \ y_3 = 0, \ y_4 = 1.$$

Tableau 11

Calcul des racines d'une équation algébrique par la méthode de Bernoulli

i	y_i	$\frac{y_i}{y_{i-1}}$	i	y_i	$\frac{y_i}{y_{i-1}}$
5	—5	—5	10	15 575	—4,992
6	25	—5	11	—77 750	—4,928
7	—125	—5	12	388 125	—4,99196
8	625	—5	13	—1 937 500	—4,991948
9	—3120	—4,992			

Calculons d'après la formule (9) les valeurs de y_i avec $i \geqslant 5$. Les valeurs obtenues sont portées sur le tableau 11. En s'arrêtant à y_{13}, on a:

$$x_1 \approx \frac{y_{13}}{y_{12}} = -\frac{1\,937\,500}{388\,125} = -4{,}991948.$$

Il en résulte, compte tenu de y_{12}, qu'on peut poser approximativement:

$$x_1 = -4{,}99195.$$

Notons en conclusion que ces dernières années sont apparues de nouvelles méthodes à schémas de calcul commodes (celles de Lin, de N. V. Pulover, etc.) [10].

BIBLIOGRAPHIE

1. *A. Kurosh.* Cours d'algèbre supérieure. Editions MIR, Moscou, 1971.
2. *G. Chapiro.* Algèbre supérieure. 4e éd., GUPI, Moscou, 1938, chapitres III, VI.
3. *D. Grave.* Eléments d'algèbre supérieure. Kiev, 1914, chapitre X.
4. *B. Fouks, B. Chabat.* Fonctions des variables complexes. Gostekhizdat, Moscou-Léningrad, 1949, chapitre VII.
5. *A. Krylov.* Conférences sur les calculs approchés. 2e éd., Editions de l'Académie des Sciences de l'U.R.S.S., chapitre II.
6. *J. B. Scarborough.* Numerical mathematical analysis. John Hopkins, 1950, chapitre X.
7. *B. Mlodzéevski.* Résolution des équations numériques. GIZ, Moscou, 1924, chapitre IV.
8. *A. Guelfond.* Calcul des différences finies. Dunod, Paris, 1962, chapitre V.
9. *F. B. Hildebrand.* Introduction to numerical analysis. New York-Toronto-London, 1956.
10. *V. Zagouskin.* Aide-mémoire de méthodes numériques de résolution des équations algébriques et transcendantes. Physmathguiz, 1960.

CHAPITRE VI

AMÉLIORATION DE LA CONVERGENCE DES SÉRIES

§ 1. Amélioration de la convergence des séries numériques

On dit que la série

$$a_1 + a_2 + \ldots + a_n + \ldots \qquad (1)$$

converge lentement si, pour obtenir sa somme avec une précision voulue, il faut prendre un très grand nombre de ses termes. Supposons, par exemple, qu'il faut trouver la somme de la série

$$S = \frac{1}{1^2} + \frac{1}{2^2} + \ldots + \frac{1}{n^2} + \ldots \qquad (2)$$

à 10^{-6} près. L'estimation du n-ième reste de la série s'écrit

$$R_n < \int_n^{\infty} \frac{dx}{x^2} = \frac{1}{n}.$$

Par conséquent, la précision imposée ne peut être assurée que par la somme de 1 000 000 de termes, ce qui est pratiquement impossible. Dans la recherche de la solution du problème donné, la série (2) doit donc être considérée comme lentement convergente.

Ainsi, l'obtention immédiate de la somme d'une telle série avec la précision imposée ε est en général difficile et même pratiquement impossible. C'est pourquoi les transformations des séries qui améliorent leur convergence acquièrent un intérêt particulier. Nous allons examiner la *transformation de Kummer* [3], [4] souvent utile pour la réalisation de la tâche imposée.

Soient la série convergente (1) et sa somme A. Choisissons une série convergente auxiliaire

$$b_1 + b_2 + \ldots + b_n + \ldots \qquad (b_n \neq 0) \qquad (2')$$

de somme B connue et telle qu'il existe une limite

$$\lim_{n \to \infty} \frac{a_n}{b_n} = q \neq 0. \qquad (3)$$

On a alors une égalité évidente

$$\sum_{n=1}^{\infty} a_n = q\sum_{n=1}^{\infty} b_n + \sum_{n=1}^{\infty} (a_n - qb_n)$$

ou

$$A = qB + \sum_{n=1}^{\infty} (a_n - qb_n). \tag{4}$$

En particulier, si $a_n \sim b_n$, on a $q = 1$ et

$$A = B + \sum_{n=1}^{\infty} (a_n - b_n). \tag{4'}$$

Par conséquent, la recherche de la somme de la série (1) est dans le cas général remplacée par la recherche de la somme de la série

$$\sum_{n=1}^{\infty} (a_n - qb_n). \tag{5}$$

Le reste de la série (5) $\overline{R}_N$ peut se mettre sous la forme

$$\overline{R}_N = \sum_{n=N+1}^{\infty} (a_n - qb_n) = \sum_{n=N+1}^{\infty} \left(1 - q\frac{b_n}{a_n}\right) a_n = \sum_{n=N+1}^{\infty} \varepsilon_n a_n,$$

où $\varepsilon_n = 1 - q\dfrac{b_n}{a_n} \to 0$ quand $n \to \infty$.

C'est pourquoi la série (5) converge en général plus vite que la série initiale (1). La difficulté principale de l'application de la transformation de Kummer consiste à choisir la série auxiliaire (2') convenable.

Montrons l'application de cette transformation à une série (1) aux signes positifs et dont les termes a_n sont des fonctions rationnelles d'une variable entière n

$$a_n = \frac{\alpha_0 n^p + \alpha_1 n^{p-1} + \ldots + \alpha_p}{\beta_0 n^q + \beta_1 n^{q-1} + \ldots + \beta_q} \quad (n = 1, 2, \ldots), \tag{6}$$

où p et q sont des entiers non négatifs et $\alpha_0 > 0$, $\beta_0 > 0$. Pour assurer la convergence de la série du terme général (6), il faut et il suffit que l'inégalité

$$q \geqslant p + 2$$

ait lieu.

Dans ce cas

$$a_n = O\left(\frac{1}{n^2}\right)^*$$

(au moins!).

Considérons les séries auxiliaires

$$S^{(m)} = \sum_{n=1}^{\infty} \frac{1}{n(n+1)\dots(n+m)} \qquad (m = 1, 2, \dots). \tag{7}$$

Etant donné que

$$\frac{1}{n(n+1)\dots(n+m)} = \frac{1}{m}\left[\frac{1}{n(n+1)\dots(n+m-1)} - \frac{1}{(n+1)(n+2)\dots(n+m)}\right],$$

il vient

$$S_N^{(m)} = \sum_{n=1}^{N} \frac{1}{n(n+1)\dots(n+m)} = \frac{1}{m}\left[\frac{1}{1\cdot 2\dots m} - \frac{1}{(N+1)(N+2)\dots(N+m)}\right].$$

Par conséquent,

$$S^{(m)} = \lim_{N\to\infty} S_N^{(m)} = \frac{1}{mm!}. \tag{8}$$

En utilisant l'idée de Stirling, le terme général de la série défini par la formule (6) est mis sous la forme d'une somme finie des factorielles inverses

$$a_n = \frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots + \frac{A_m}{n(n+1)\dots(n+m)} + a_n^{(m)},$$

où $A_1, A_2, \dots, A_m$ sont les coefficients indéterminés et $a_n^{(m)}$ est le reste. Choisissons les coefficients A_i ($i = 1, 2, \dots, m$) de sorte que

$$a_n^{(m)} = O\left(\frac{1}{n^{2+m}}\right).$$

---

* On dit que a_n est un infiniment petit au moins d'ordre m par rapport à $\frac{1}{n}$:

$$a_n = O\left(\frac{1}{n^m}\right),$$

si

$$\lim_{n\to\infty} \frac{a_n}{\left(\frac{1}{n}\right)^m} = c \neq \infty.$$

Si en outre $c \neq 0$, a_n est un infiniment petit exactement d'ordre m par rapport à $\frac{1}{n}$.

A cet effet il suffit de définir successivement les coefficients A_i d'après les formules

$$\left.\begin{aligned} A_1 &= \lim_{n\to\infty} n(n+1)\,a_n, \\ A_2 &= \lim_{n\to\infty}\left[a_n - \frac{A_1}{n(n+1)}\right] n(n+1)(n+2) \\ &\dots\dots\dots\dots\dots\dots\dots\dots \\ A_m &= \lim_{n\to\infty}\left[a_n - \sum_{i=1}^{m-1}\frac{A_i}{n(n+1)\dots(n+i)}\right] n(n+1)\dots(n+m). \end{aligned}\right\} \quad (9)$$

Conformément au schéma général, on prend pour série auxiliaire (2):

$$B = \sum_{n=1}^{\infty} b_n =$$

$$= \sum_{n=1}^{\infty}\left[\frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots + \frac{A_m}{n(n+1)\dots(n+m)}\right] =$$

$$= A_1 S^{(1)} + A_2 S^{(2)} + \dots + A_m S^{(m)} = \frac{A_1}{1\cdot 1!} + \frac{A_2}{2\cdot 2!} + \dots + \frac{A_m}{m\cdot m!}, \quad (10)$$

par ailleurs, il est clair que

$$\lim_{n\to\infty}\frac{a_n}{b_n} = 1$$

et

$$S = \sum_{n=1}^{\infty} a_n = B + \sum_{n=1}^{\infty} a_n^{(m)}. \quad (11)$$

La convergence rapide de la série auxiliaire $\sum_{n=1}^{\infty} a_n^{(m)}$ ne se manifestant en général que lorsque n est suffisamment grand, il est pratiquement commode de ne réaliser la transformation considérée qu'en partant d'un certain terme a_{p+1} de la série. En posant

$$S = \sum_{n=1}^{p} a_n + \sum_{n=p+1}^{\infty} a_n = S_p + \sum_{n=p+1}^{\infty} a_n,$$

on a:

$$S = S_p +$$

$$+ \sum_{n=p+1}^{\infty}\left[\frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots + \frac{A_m}{n(n+1)\dots(n+m)} + a_n^{(m)}\right] =$$

$$= S_p + A_1 \sum_{n=p+1}^{\infty}\left(\frac{1}{n} - \frac{1}{n+1}\right) +$$

$$+\frac{A_2}{2}\sum_{n=p+1}^{\infty}\left[\frac{1}{n(n+1)}-\frac{1}{(n+1)(n+2)}\right]+\ldots$$

$$\ldots+\frac{A_m}{m}\sum_{n=p+1}^{\infty}\left[\frac{1}{n(n+1)\ldots(n+m-1)}-\frac{1}{(n+1)\ldots(n+m)}\right]+$$

$$+\sum_{n=p+1}^{\infty}a_n^{(m)}=S_p+A_1\cdot\frac{1}{p+1}+\frac{A_2}{2}\cdot\frac{1}{(p+1)(p+2)}+\ldots$$

$$\ldots+\frac{A_m}{m}\cdot\frac{1}{(p+1)\ldots(p+m)}+\sum_{n=p+1}^{\infty}a_n^{(m)}.$$

En particulier, pour $m\to\infty$ et tenant compte de $a_n^{(m)}\rightrightarrows 0$, on obtient le *développement de Stirling*

$$\sum_{n=1}^{\infty}a_n=\sum_{n=1}^{p}a_n+A_1\cdot\frac{1}{p+1}+\frac{A_2}{2}\cdot\frac{1}{(p+1)(p+2)}+\ldots$$

$$\ldots+\frac{A_m}{m}\cdot\frac{1}{(p+1)(p+2)\ldots(p+m)}+\ldots$$

Exemple. Trouver la somme de la série

$$S=\sum_{n=1}^{\infty}\frac{1}{n^2+1}\tag{12}$$

à 0,001 près.

Solution. Posons:

$$\frac{1}{n^2+1}=\frac{A_1}{n(n+1)}+\frac{A_2}{n(n+1)(n+2)}+a_n^{(2)}.$$

On a:

$$A_1=\lim_{n\to\infty}\frac{n(n+1)}{n^2+1}=1;$$

$$A_2=\lim_{n\to\infty}\left[\frac{1}{n^2+1}-\frac{1}{n(n+1)}\right]n(n+1)(n+2)=\lim_{n\to\infty}\frac{(n-1)(n+2)}{n^2+1}=1.$$

Par conséquent,

$$a_n^{(2)}=\frac{1}{n^2+1}-\frac{1}{n(n+1)}-\frac{1}{n(n+1)(n+2)}=$$

$$=\frac{n^3+3n^2+2n-n^3-2n^2-n-2-n^2-1}{n(n+1)(n+2)(n^2+1)}=\frac{n-3}{n(n+1)(n+2)(n^2+1)}.$$

Les formules (10) et (11) entraînent

$$S=\frac{1}{1\cdot 1\,!}+\frac{1}{2\cdot 2\,!}+\sum_{n=1}^{\infty}\frac{n-3}{n(n+1)(n+2)(n^2+1)}.\tag{12'}$$

Puisque avec $n \geqslant 3$ on a

$$\frac{n-3}{n(n+1)(n+2)(n^2+1)} \leqslant \frac{1}{n^4},$$

il vient

$$\rho_N = \sum_{n=N+1}^{\infty} \frac{n-3}{n(n+1)(n+2)(n^2+1)} < \int_N^{\infty} \frac{dx}{x^4} = \frac{1}{3N^3} < \frac{1}{2} \cdot 0{,}001.$$

Il en résulte que le nombre suffisant de termes de la somme (12′) est $N = 10$; en outre, ces termes doivent être calculés avec quatre décimales au sens strict. Ainsi, on a:

$$S \approx 1{,}25 + (-0{,}1667) + (-0{,}0083) + 0 + 0{,}0005 + 0{,}0004 + \\ + 0{,}0002 + 0{,}0002 + 0{,}0001 + 0{,}0001 + 0{,}0001 = 1{,}0766,$$

de plus, en retenant que la somme des quatre premiers termes est exacte, on obtient pour l'erreur absolue l'estimation

$$\Delta < \frac{1}{3} \cdot 10^{-3} + 7 \cdot \frac{1}{2} \cdot 10^{-4} < 0{,}7 \cdot 10^{-3}.$$

On en tire, en arrondissant, la quantité

$$S \approx 1{,}077$$

avec une borne d'erreur absolue

$$\bar{\Delta} = 0{,}7 \cdot 10^{-3} + 0{,}4 \cdot 10^{-3} = 1{,}1 \cdot 10^{-3}.$$

Constatons que l'estimation du reste de la série (12) s'écrit

$$R_N < \int_N^{\infty} \frac{dx}{x^2+1} < \int_N^{\infty} \frac{dx}{x^2} = \frac{1}{N} \leqslant \frac{1}{2} \cdot 0{,}001.$$

D'où $N \geqslant 2000$, donc pour obtenir la même précision sans transformation il faut environ 2000 termes de la série.

R e m a r q u e. Pour calculer la somme approchée de la série (1) du terme général (6) on peut faire également appel aux séries

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}; \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}; \quad \sum_{n=1}^{\infty} \frac{1}{n^6} = \frac{\pi^6}{945}, \text{ etc.}$$

En général

$$\sum_{n=1}^{\infty} \frac{1}{n^{2p}} = \frac{(-1)^{p-1}}{2} \cdot \frac{B_{2p}(2\pi)^{2p}}{(2p)!},$$

où B_n ($n = 1, 2, \ldots$) sont les nombres de Bernoulli [5], [6] définis par la formule symbolique

$$(B + 1)^n - B^n = 0,$$

dans laquelle après le développement suivant le binôme de Newton on pose $B^n = B_n$. On a en particulier

$$B_2 = \frac{1}{6}; \quad B_4 = -\frac{1}{30}; \quad B_6 = \frac{1}{42}; \quad B_8 = -\frac{1}{30}; \quad B_{10} = \frac{5}{66}$$

(cf. chapitre XVI, § 11).

§ 2. Amélioration de la convergence des séries entières par la méthode d'Euler-Abel

Considérons la série entière convergente

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \tag{1}$$

où $f(x)$ est la somme de la série.

Supposons que le rayon de convergence R de la série (1) soit fini et différent de zéro. Sans porter atteinte à la généralité du raisonnement on peut considérer que $R = 1$ *.

Mettons la série (1) sous la forme suivante:

$$f(x) = a_0 + x\varphi(x), \tag{2}$$

où

$$\varphi(x) = \sum_{n=1}^{\infty} a_n x^{n-1} = \sum_{n=0}^{\infty} a_{n+1} x^n. \tag{3}$$

En multipliant les deux membres de l'égalité (3) par le binôme $1 - x$, on obtient:

$$(1 - x)\varphi(x) = \sum_{n=0}^{\infty} a_{n+1} x^n - \sum_{n=0}^{\infty} a_{n+1} x^{n+1}. \tag{4}$$

En posant dans la deuxième somme $n + 1 = m$ et en tenant compte de ce que la somme ne dépend pas de la désignation de l'indice de sommation, on a:

$$\sum_{n=0}^{\infty} a_{n+1} x^{n+1} = \sum_{m=1}^{\infty} a_m x^m = \sum_{n=1}^{\infty} a_n x^n.$$

---

* En effet, si $0 < R < \infty$ et $R \neq 1$, en posant $t = \frac{x}{R}$ on obtient une série entière par rapport à la variable t de rayon de convergence $\rho = 1$.

Donc

$$(1-x)\,\varphi(x)=\sum_{n=0}^{\infty} a_{n+1}x^n-\sum_{n=1}^{\infty} a_n x^n=$$
$$=a_0+\sum_{n=0}^{\infty}(a_{n+1}-a_n)\,x^n=a_0+\sum_{n=0}^{\infty}\Delta a_n x^n,$$

où

$$\Delta a_n=a_{n+1}-a_n \qquad (n=0,\ 1,\ 2,\ \ldots)$$

sont les *différences premières* des coefficients a_n (pour plus de détail sur les différences finies cf. chapitre XIV, § 1). Par conséquent, les formules (3) et (4) permettent de déduire:

$$\varphi(x)=\sum_{n=0}^{\infty} a_{n+1}x^n=\frac{a_0}{1-x}+\frac{1}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n$$

et donc

$$f(x)=a_0+\frac{a_0 x}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n=\frac{a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n,$$

soit

$$\sum_{n=0}^{\infty} a_n x^n=\frac{a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n x^n. \qquad (5)$$

La transformation considérée de la série entière s'appelle *transformation d'Euler-Abel*. En appliquant d'une façon analogue la transformation d'Euler-Abel à la série entière $\sum_{n=0}^{\infty}\Delta a_n x^n$, on trouve:

$$\sum_{n=0}^{\infty}\Delta a_n x^n=\frac{\Delta a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta^2 a_n x^n,$$

où

$$\Delta^2 a_n = \Delta\,(\Delta a_n) = \Delta a_{n+1} - \Delta a_n$$

sont les *différences secondes* des coefficients a_n. On en tire en vertu de la formule (5):

$$\sum_{n=0}^{\infty} a_n x^n=\frac{a_0}{1-x}+\frac{x}{1-x}\left(\frac{\Delta a_0}{1-x}+\frac{x}{1-x}\sum_{n=0}^{\infty}\Delta^2 a_n x^n\right)=$$
$$=\frac{a_0}{1-x}+\frac{x\Delta a_0}{(1-x)^2}+\left(\frac{x}{1-x}\right)^2\sum_{n=0}^{\infty}\Delta^2 a_n x^n.$$

En reprenant successivement p fois la transformation d'Euler-Abel on a

$$\sum_{n=0}^{\infty} a_n x^n = \frac{a_0}{1-x} + \frac{x\Delta a_0}{(1-x)^2} + \ldots + \frac{x^{p-1}\Delta^{p-1}a_0}{(1-x)^p} + \left(\frac{x}{1-x}\right)^p \sum_{n=0}^{\infty} \Delta^p a_n x^n,$$

où

$$\Delta^p a_n = \Delta^{p-1}a_{n+1} - \Delta^{p-1}a_n \qquad (n=0,\ 1,\ 2,\ \ldots)$$

sont les *différences d'ordre* p des coefficients a_n et $\Delta^k a_0$ ($k = 0, 1, 2, \ldots$) les différences finies consécutives des coefficients a_n avec $n = 0$. Ainsi,

$$f(x) = \sum_{k=0}^{p-1} \Delta^k a_0 \frac{x^k}{(1-x)^{k+1}} = \left(\frac{x}{1-x}\right)^p \sum_{n=0}^{\infty} \Delta^p a_n x^n, \tag{6}$$

où l'on a posé $\Delta^0 a_0 = a_0$. Si l'ordre de décroissance des différences finies $\Delta^p a_n$ pour $n \to \infty$ est supérieur à celui des coefficients a_n, il est plus avantageux d'employer la formule (6). Cette condition se présente assez souvent. Par exemple, si $a_n = \frac{1}{n}$, on obtient:

$$\Delta a_n = \frac{1}{n+1} - \frac{1}{n} = -\frac{1}{n(n+1)};$$

ici, quand $n \to \infty$, Δa_n diminue plus vite que a_n.

En particulier, si $a_n = P(n)$, où $P(n)$ est un polynôme entier de degré $p - 1$, la formule (6) donne sous une forme finie la somme de la série

$$\sum_{n=0}^{\infty} P(n)\, x^n = \sum_{k=0}^{p-1} \Delta^k P(0) \frac{x^k}{(1-x)^{k+1}} \qquad (|x| < 1), \tag{7}$$

du fait que $\Delta^p P(n) = 0$.

La formule (6) n'a pas de sens avec $x = 1$. Pour ce cas, la transformation d'Euler-Abel peut être modifiée. En posant $x = -t$, on a:

$$f(x) = \sum_{n=0}^{\infty} a_n(-t)^n = \sum_{n=0}^{\infty} (-1)^n a_n t^n =$$

$$= \sum_{k=0}^{p-1} \Delta^k [(-1)^n a_n]_{n=0} \frac{t^k}{(1-t)^{k+1}} + \left(\frac{t}{1-t}\right)^p \sum_{n=0}^{\infty} \Delta^p [(-1)^n a_n]\, t^n.$$

En revenant à l'ancienne variable, on obtient:

$$f(x)=\sum_{k=0}^{p-1}(-1)^k\Delta^k[(-1)^n a_n]_{n=0}\frac{x^k}{(1+x)^k}+$$
$$+\left(\frac{x}{1+x}\right)^p\sum_{n=0}^{\infty}(-1)^{n+p}\Delta^p[(-1)^n a_n]\,x^n. \quad (8)$$

La formule (8) a également un sens pour $x = 1$.

Exemple 1. Calculer à 0,001 près la somme de la série

$$f(x)=\sum_{n=0}^{\infty}\frac{x^n}{(n+1)(n+2)} \quad (9)$$

pour $x = -1$.

Solution. Appliquons deux fois la transformation d'Euler ($p = 2$). On a:

$$a_n=\frac{1}{(n+1)(n+2)};$$
$$\Delta a_n=a_{n+1}-a_n=\frac{1}{(n+2)(n+3)}-\frac{1}{(n+1)(n+2)}=$$
$$=-\frac{2}{(n+1)(n+2)(n+3)};$$
$$\Delta^2 a_n=\Delta a_{n+1}-\Delta a_n=-\frac{2}{(n+2)(n+3)(n+4)}+\frac{2}{(n+1)(n+2)(n+3)}=$$
$$=\frac{6}{(n+1)(n+2)(n+3)(n+4)}.$$

Par conséquent,

$$a_0=\frac{1}{1\cdot2};\quad \Delta a_0=-\frac{2}{1\cdot2\cdot3}.$$

La formule (6) entraîne:

$$f(-1)=\frac{1}{1\cdot2}\cdot\frac{1}{2}+\frac{2}{1\cdot2\cdot3}\cdot\frac{1}{4}+$$
$$+\left(-\frac{1}{2}\right)^2\sum_{n=0}^{\infty}\frac{6}{(n+1)(n+2)(n+3)(n+4)}(-1)^n=$$
$$=\frac{1}{4}+\frac{1}{12}+\frac{3}{2}\cdot\frac{1}{24}-\frac{3}{2}\cdot\frac{1}{120}+\frac{3}{2}\cdot\frac{1}{360}-\frac{3}{2}\cdot\frac{1}{840}+$$
$$+\frac{3}{2}\cdot\frac{1}{1680}-\frac{3}{2}\cdot\frac{1}{3024}+\frac{3}{2}\cdot\frac{1}{5040}-\dots \quad (10)$$

La série (10) est une série alternée à termes décroissants en module. Par conséquent, si nous nous arrêtons au terme

$$\frac{3}{2}\cdot\frac{1}{3024}=\frac{1}{2016},$$

le reste de la série R ne sera pas supérieur en module au premier terme négligé:

$$|R| < \frac{3}{2}\cdot\frac{1}{5040} = \frac{1}{3360} < 3\cdot 10^{-4}.$$

Ainsi, si l'on prend deux chiffres de réserve, on a:

$$f(-1) = 0,25000 + 0,08333 + 0,06250 - 0,01250 + 0,00417 - \\ - 0,00179 + 0,00089 - 0,00050 = 0,38610$$

avec une erreur absolue

$$\Delta < 5\cdot\frac{1}{2}\cdot 10^{-5} + 3\cdot 10^{-4} < 4\cdot 10^{-4}.$$

En arrondissant le nombre obtenu à trois chiffres, on obtient la valeur approchée de $f(-1) = 0,386$ avec une borne d'erreur absolue

$$\Delta < 4\cdot 10^{-4} + 1\cdot 10^{-4} = \frac{1}{2}\cdot 10^{-3}.$$

La valeur exacte de la somme est:

$$f(-1) = 2\ln 2 - 1 = 0,38630\ldots$$

Notons que si l'on applique la série (9) pour calculer directement $f(-1)$, la précision imposée ne peut s'obtenir qu'en prenant à peu près quarante cinq termes de cette série.

Exemple 2. Trouver la somme de la série

$$S(x) = \sum_{n=0}^{\infty} (n^2 + n + 1)\, x^n.$$

Solution. Soit

$$P(n) = n^2 + n + 1.$$

Composons le tableau 12.

Tableau 12

Tableau des différences finies

n	$P(n)$	$\Delta P(n)$	$\Delta^2 P(n)$
0	1	2	2
1	3	4	
2	7		

La formule (7) amène:

$$S(x) = \frac{1}{1-x} + \frac{2x}{(1-x)^2} + \frac{2x^2}{(1-x)^3}$$

avec $|x| < 1$.

§ 3. Estimations des coefficients de Fourier

On appelle série trigonométrique ou série de Fourier d'une fonction donnée $f(x)$ $(-\pi < x < \pi)$ * la série

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx), \tag{1}$$

dont les coefficients a_n, b_n (*coefficients de Fourier* de la fonction $f(x)$) se calculent d'après les formules

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx \quad (n = 0, 1, \ldots), \tag{2}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx \quad (n = 1, 2, \ldots). \tag{2'}$$

La condition suffisante de l'existence d'une série de Fourier de $f(x)$ est l'intégrabilité de cette fonction sur le segment $[-\pi, \pi]$. Dans ce cas les coefficients de Fourier (2) et (2′) ont des valeurs finies déterminées.

Il se peut que la série de Fourier obtenue diverge ou converge vers une autre fonction. Donnons sans démonstration les conditions suffisantes de convergence de la série trigonométrique vers la fonction $f(x)$ en tout point de continuité de cette dernière [1], [7].

Théorème de convergence. *Si la fonction $f(x)$ est continue par morceaux et dérivable par morceaux sur le segment $[-\pi, \pi]$, sa série de Fourier converge sur l'axe numérique tout entier et sa somme $S(x)$ est une fonction 2π-périodique égale à*

$$S(x_0) = \frac{f(x_0 - 0) + f(x_0 + 0)}{2} \tag{3}$$

en tout point $x_0 \in (-\pi, \pi)$ et $S(\pm\pi) = 2^{-1} [f(-\pi + 0) + f(\pi - 0)]$.

En particulier, $S(x_0) = f(x_0)$ si en $x = x_0$ la fonction est continue, c'est-à-dire si $f(x_0 - 0) = f(x_0 + 0) = f(x_0)$.

Si, en outre, la fonction $f(x)$ est 2π-périodique, sa série de Fourier converge pour toute valeur de x_0 et sa somme est donnée par (3).

Si les conditions du théorème de convergence sont respectées, il est clair que $a_n \to 0$ et $b_n \to 0$ quand $n \to 0$. Nous donnerons des estimations plus précises des coefficients de Fourier qui s'obtiennent

---

* Pour abréger les énoncés nous considérons la fonction définie sur le segment $[-\pi, \pi]$. Le cas général d'une fonction $\varphi(t)$ définie sur le segment $[a, b]$ peut être ramené au cas considéré à l'aide de la substitution linéaire

$$t = \frac{b+a}{2} + \frac{b-a}{2\pi} x.$$

en soumettant le comportement de la fonction $f(x)$ à certaines restrictions.

D é f i n i t i o n. *On dit qu'une fonction $f(x)$ définie sur le segment $[-\pi, \pi]$ appartient à la classe de périodicité $\widetilde{C}^{(m)}$ si*

1) *$f(x)$ et ses dérivées jusqu'à l'ordre m-ième y compris sont continues sur le segment $[-\pi, \pi]$;*

2) $f^{(k)}(-\pi + 0) = f^{(k)}(\pi - 0)$ *pour* $k = 0, 1, 2, \ldots, m$, *c'est-à-dire les valeurs de la fonction $f(x)$ et de ses m premières dérivées coïncident aux extrémités du segment $[-\pi, \pi]$.*

Il s'ensuit des conditions 1) et 2) que le prolongement périodique de la fonction $f(x)$ appartient à la classe $C^{(m)}(-\infty, +\infty)$.

L e m m e. *Si la fonction $f(x)$ appartient à la classe de périodicité $\widetilde{C}^{(m)}$ sur le segment $[-\pi, \pi]$* (en abrégé, $f(x) \in \widetilde{C}^{(m)}[-\pi, \pi]$), *ses coefficients de Fourier a_n et b_n sont pour $n \to \infty$ des infiniment petits d'ordre supérieur à m par rapport à $\frac{1}{n}$, c'est-à-dire*

$$a_n = o\left(\frac{1}{n^m}\right); \quad b_n = o\left(\frac{1}{n^m}\right)^*.$$

D é m o n s t r a t i o n. Intégrons par parties m fois les deuxièmes membres des égalités suivantes:

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx \quad (n = 0, 1, \ldots), \tag{4}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx \quad (n = 1, 2, \ldots) \tag{4'}$$

Si l'on pose $u = f(x)$ et $dv = \cos nx \, dx$, on obtient $du = f'(x) \, dx$ et $v = \frac{1}{n} \sin nx$. Par conséquent, la formule d'intégration par parties conduit à:

$$a_n = \frac{1}{\pi}\left[\frac{1}{n} f(x) \sin nx\right]_{-\pi}^{\pi} - \frac{1}{\pi n} \int_{-\pi}^{\pi} f'(x) \sin nx \, dx =$$

$$= \frac{1}{\pi n} \int_{-\pi}^{\pi} f'(x) \cos\left(\frac{\pi}{2} + nx\right) dx.$$

---

* La notation $a_n = o\left(\frac{1}{n^m}\right)$ signifie que $\lim_{n \to \infty} \left(\frac{a_n}{\frac{1}{n^m}}\right) = 0$.

14—01072

En procédant encore une fois à l'intégration par parties et en prenant en considération que $f'(-\pi) = f'(\pi)$, on obtient:

$$a_n = \frac{1}{\pi n}\left\{\left[\frac{1}{n} f'(x)\sin\left(\frac{\pi}{2} + nx\right)\right|_{-\pi}^{\pi} + \right.$$

$$\left. + \frac{1}{n}\int_{-\pi}^{\pi} f''(x)\cos\left(\frac{\pi}{2}\cdot 2 + nx\right)dx\right]\right\} =$$

$$= \frac{1}{\pi n^2}\int_{-\pi}^{\pi} f''(x)\cos\left(\frac{\pi}{2}\cdot 2 + nx\right)dx,$$

etc.

Après m intégrations par parties, les formules (4) et (4′) entraînent

$$a_n = \frac{1}{\pi n^m}\int_{-\pi}^{\pi} f^{(m)}(x)\cos\left(\frac{\pi}{2}\cdot m + nx\right)dx.$$

D'une façon analogue

$$b_n = \frac{1}{\pi n^m}\int_{-\pi}^{\pi} f^{(m)}(x)\sin\left(\frac{\pi}{2}\cdot m + nx\right)dx.$$

Les intégrales

$$\varepsilon_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f^{(m)}(x)\cos\left(\frac{\pi}{2}\cdot m + nx\right)dx$$

et

$$\varepsilon_n' = \frac{1}{\pi}\int_{-\pi}^{\pi} f^{(m)}(x)\sin\left(\frac{\pi}{2}\cdot m + nx\right)dx$$

sont au signe près les coefficients de Fourier de la fonction $f^{(m)}(x)$ continue par hypothèse. On sait qu'indépendamment de la convergence ou de la divergence de la série de Fourier, les coefficients de Fourier d'une fonction continue tendent vers zéro lorsque leur rang tend vers l'infini *. Il en résulte que

$$\varepsilon_n \to 0 \text{ et } \varepsilon_n' \to 0 \text{ quand } n \to \infty.$$

* Il en est ainsi du fait que toute fonction continue par morceaux $f(x)$ à coefficients de Fourier a_n et b_n $(n = 0, 1, 2, \ldots)$ vérifie l'*inégalité de Bessel* [7]

$$\frac{a_0^2}{2} + \sum_{n=1}^{\infty}(a_n^2 + b_n^2) \leqslant \frac{1}{\pi}\int_{-\pi}^{\pi} f^2(x)\,dx.$$

Par conséquent, la série $\sum_{n=1}^{\infty}(a_n^2 + b_n^2)$ converge et $a_n \to 0$; $b_n \to 0$ quand $n \to \infty$.

Or

$$a_n = \frac{\varepsilon_n}{n^m} \quad \text{et} \quad b_n = \frac{\varepsilon'_n}{n^m},$$

donc les coefficients de Fourier a_n et b_n de la fonction $f(x)$ sont des infiniment petits d'ordre supérieur par rapport à $\frac{1}{n^m}$:

$$a_n = o\left(\frac{1}{n^m}\right), \quad b_n = o\left(\frac{1}{n^m}\right).$$

Ce résultat a été utilisé par A. Krylov qui l'a mis à la base de sa méthode d'amélioration de la convergence des séries de Fourier.

Remarque. Si $f^{(m)}(x)$ satisfait aux conditions du théorème de convergence, on montre sans peine que

$$\varepsilon_n = O\left(\frac{1}{n}\right) \quad \text{et} \quad \varepsilon'_n = O\left(\frac{1}{n}\right).$$

Dans ce cas l'estimation des coefficients de Fourier est bien meilleure :

$$a_n = O\left(\frac{1}{n^{m+1}}\right) \quad \text{et} \quad b_n = O\left(\frac{1}{n^{m+1}}\right).$$

§ 4. Amélioration de la convergence des séries de Fourier par la méthode de A. Krylov

Soient une fonction $f(x)$ continue par morceaux sur le segment $[-\pi, \pi]$ et ses dérivées continues par morceaux $f^{(i)}(x)$ $(i = 1, 2, \ldots, m)$ jusqu'à l'ordre m-ième y compris. En vertu du théorème de la convergence (§ 3), la fonction $f(x)$ peut être représentée en tout point de continuité par la série trigonométrique de Fourier

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx), \tag{1}$$

où a_n et b_n sont les coefficients de Fourier définis par les formules (2) et (2′) du § 3. Dans le cas général, les coefficients a_n et b_n tendent vers zéro lentement et il est pratiquement difficile d'utiliser la série (1) ; il est d'autant plus inadmissible de dériver la série (1) terme à terme, opération quelquefois imposée par la résolution de certains problèmes, en particulier par l'application de la méthode de Fourier.

La méthode de Krylov [8] consiste à extraire de la fonction $f(x)$ une fonction élémentaire $g(x)$ (qui est en général une fonction polynomiale par morceaux) de mêmes discontinuités que la fonction $f(x)$, les dérivées $g^{(i)}(x)$ $(i = 1, 2, \ldots, m)$ jusqu'à l'ordre m-ième y compris possédant les mêmes discontinuités que les dérivées correspondantes $f^{(i)}(x)$ de la fonction considérée $f(x)$ et de plus

la fonction $g(x)$ étant telle que

$$f^{(i)}(-\pi+0)-g^{(i)}(-\pi+0)=$$
$$=f^{(i)}(\pi-0)-g^{(i)}(\pi-0) \quad (i=0, 1, 2, \ldots, m).$$

Dans ce cas, la différence

$$\varphi(x) = f(x) - g(x)$$

appartient à la classe de périodicité $\tilde{C}^{(m)}$.

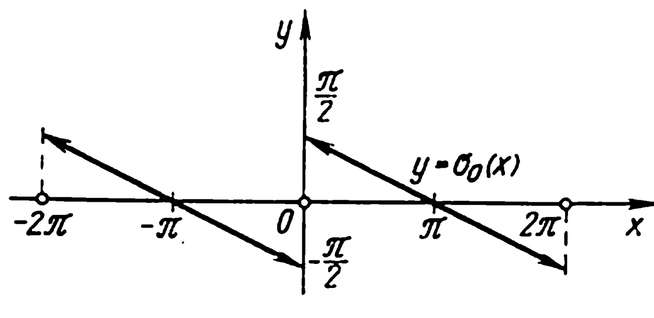


Fig. 46.

En désignant par α_n et β_n $(n = 0, 1, 2, \ldots)$ les coefficients de Fourier de la fonction $\varphi(x)$, on aura:

$$f(x)=g(x)+\left[\frac{\alpha_0}{2}+\sum_{n=1}^{\infty}(\alpha_n \cos nx+\beta_n \sin nx)\right], \tag{2}$$

où α_n et β_n sont des infiniment petits d'ordre supérieur à m par rapport à $\frac{1}{n}$, quand $n \to \infty$, c'est-à-dire en général la série (2) converge rapidement. Cette série peut être dérivée terme à terme au moins $m-2$ fois.

Montrons comment en pratique on construit la fonction auxiliaire $g(x)$ à partir de la fonction donnée $f(x)$ [9]. Construisons à cet effet par la méthode de récurrence sur le segment $[-2\pi, 2\pi]$ la suite des fonctions $\sigma_0(x)$, $\sigma_1(x)$, ..., $\sigma_m(x)$, qui jouissent de la propriété suivante:

$$\sigma_k^{(k)}(+0)-\sigma_k^{(k)}(-0)=\pi \tag{3}$$

$(k = 0, 1, 2, \ldots, m)$, et de plus telles que les dérivées $\sigma_k^{(j)}(x)$ $(j = 0, 1, \ldots, k-1)$ soient continues sur le segment $[-2\pi, 2\pi]$.

La fonction $\sigma_0(x)$ est définie de la façon suivante:

$$\sigma_0(x)=\begin{cases} \dfrac{-\pi-x}{2} & \text{pour } -2\pi<x<0, \\ \dfrac{\pi-x}{2} & \text{pour } 0<x<2\pi, \\ 0 & \text{pour } x=-2\pi, 0, 2\pi. \end{cases} \tag{4}$$

Sa courbe est représentée sur la figure 46. C'est une fonction impaire, sa série de Fourier ne contient donc que les sinus des arcs mul-

tiples:

$$\sigma_0(x) = \sum_{n=1}^{\infty} b_n \sin nx,$$

où

$$b_n = \frac{2}{\pi} \int_0^{\pi} \frac{\pi - x}{2} \sin nx\, dx =$$

$$= \frac{2}{\pi} \left(-\frac{\pi - x}{2} \cdot \frac{\cos nx}{n} \Big|_0^{\pi} - \frac{1}{2\pi} \int_0^{\pi} \cos nx\, dx \right) =$$

$$= \frac{2}{\pi} \left(\frac{\pi}{2n} - \frac{1}{2n^2} \sin nx \Big|_0^{\pi} \right) = \frac{1}{n}.$$

Par suite

$$\sigma_0(x) = \frac{\sin x}{1} + \frac{\sin 2x}{2} + \ldots + \frac{\sin nx}{n} + \ldots \tag{5}$$

Il est évident que la fonction $\sigma_0(x)$ comporte une discontinuité au point $x = 0$ avec un saut égal à π:

$$\sigma_0(+0) - \sigma_0(-0) = \frac{\pi}{2} - \left(-\frac{\pi}{2} \right) = \pi.$$

C'est pourquoi la fonction

$$\psi(x) = \sigma_0(x - x_0) \qquad (-\pi \leqslant x \leqslant \pi;\ -\pi \leqslant x_0 \leqslant \pi)$$

fait en x_0 le même saut que la fonction $\sigma_0(x)$:

$$\psi(x_0 + 0) - \psi(x_0 - 0) = \pi,$$

le point de discontinuité étant unique sur le segment $[-\pi, \pi]$.

Définissons la fonction $\sigma_1(x)$ par la formule

$$\sigma_1(x) = c_1 + \int_0^x \sigma_0(x)\, dx, \tag{6}$$

où c_1 est une certaine constante.

Intégrons terme à terme la série (5) pour obtenir:

$$\sigma_1(x) = c_1 + \int_0^x \sum_{n=1}^{\infty} \frac{\sin nx}{n}\, dx = c_1 + \sum_{n=1}^{\infty} \frac{1}{n^2} - \sum_{n=1}^{\infty} \frac{\cos nx}{n^2}. \tag{7}$$

Choisissons la constante c_1 de façon que le terme constant de la série (7) soit nul

$$c_1 + \sum_{n=1}^{\infty} \frac{1}{n^2} = 0.$$

Il vient

$$c_1 = -\sum_{n=1}^{\infty} \frac{1}{n^2}.$$

La série $\sum_{n=1}^{\infty} \frac{1}{n^2}$ est égale évidemment au terme constant de la série de Fourier de la fonction $\int_0^x \sigma_0(x)\,dx$. On en tire en utilisant

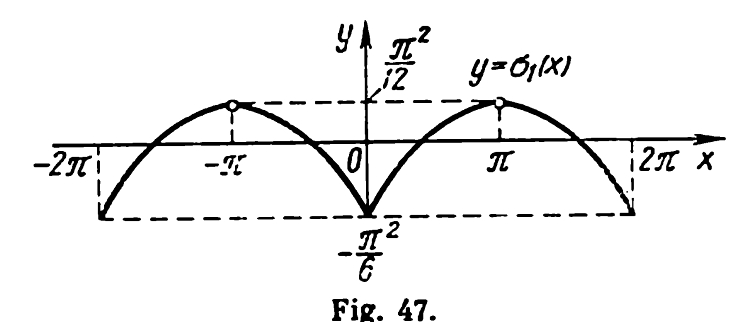


Fig. 47.

la formule (4):

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{\pi}\int_0^{\pi} dx \int_0^{x} \sigma_0(x)\,dx = \frac{1}{\pi}\int_0^{\pi}\left[\frac{\pi^2}{4} - \frac{(\pi-x)^2}{4}\right]dx =$$

$$= \frac{1}{\pi}\left(\frac{\pi^3}{4} - \frac{\pi^3}{12}\right) = \frac{\pi^2}{6}.$$

C'est pourquoi

$$c_1 = -\frac{\pi^2}{6}.$$

Donc

$$\sigma_1(x) = -\sum_{n=1}^{\infty} \frac{\cos nx}{n^2}, \tag{8}$$

en outre,

$$\sigma_1(x) = \begin{cases} \int_0^x \frac{\pi - x}{2}\,dx - \frac{\pi^2}{6} = \frac{\pi^2}{12} - \frac{(\pi - x)^2}{4} & \text{avec } 0 \leqslant x \leqslant 2\pi, \\ -\int_0^x \frac{\pi + x}{2}\,dx - \frac{\pi^2}{6} = \frac{\pi^2}{12} - \frac{(\pi + x)^2}{4} & \text{avec } -2\pi \leqslant x \leqslant 0. \end{cases}$$

La courbe de la fonction $\sigma_1(x)$ est représentée sur la figure 47. La fonction $\sigma_1(x)$ est continue sur le segment $[-2\pi, 2\pi]$, alors que sa dérivée $\sigma_1'(x) = \sigma_0(x)$ admet une discontinuité en $x = 0$; de plus

$$\sigma_1'(+0) - \sigma_0'(-0) = \pi.$$

On définit de la même façon les fonctions suivantes

$$\sigma_2(x) = \int_0^x \sigma_1(x)\,dx + c_2;$$

$$\sigma_3(x) = \int_0^x \sigma_2(x)\,dx + c_3;$$

. . . .

$$\sigma_m(x) = \int_0^x \sigma_{m-1}(x)\,dx + c_m,$$

où les constantes arbitraires $c_1, c_2, \ldots, c_m$ sont choisies de façon que le terme constant de la série de Fourier correspondante soit nul, c'est-à-dire que les constantes c_k $(k = 1, 2, \ldots, m)$ s'obtiennent successivement d'après les conditions

$$\int_0^\pi \left[c_k + \int_0^x \sigma_{k-1}(x)\,dx\right] dx = 0.$$

Les fonctions $\sigma_k(x)$ $(k = 1, 2, \ldots, m)$ et toutes les dérivées jusqu'à l'ordre $(k-1)$ y compris sont continues sur le segment $[-2\pi, 2\pi]$. Par ailleurs, comme $\sigma_k^{(k)}(x) = \sigma_0(x)$, il vient

$$\sigma_k^{(k)}(+0) - \sigma_k^{(k)}(-0) = \pi \quad (k = 1, 2, \ldots, m),$$

donc la dérivée k-ième de la fonction $\sigma_k(x)$ a une discontinuité en $x = 0$ avec un saut égal à π. Il en résulte que la fonction $\psi_k(x) = \sigma_k(x - x_0)$ $(-\pi \leqslant x \leqslant \pi)$, obtenue par translation de la fonction $\sigma_k(x)$, n'a une discontinuité que de la dérivée k-ième au point $x = x_0$:

$$\psi_k^{(k)}(x_0 + 0) - \psi_k^{(k)}(x_0 - 0) = \pi.$$

Soient maintenant

$x_1^{(0)}, x_2^{(0)}, \ldots, x_{k_0}^{(0)}$ les points de discontinuité de $f(x)$;

$x_1^{(1)}, x_2^{(1)}, \ldots, x_{k_1}^{(1)}$ les points de discontinuité de $f'(x)$;

. . . .

$x_1^{(m)}, x_2^{(m)}, \ldots, x_{k_m}^{(m)}$ les points de discontinuité de $f^{(m)}(x)$,

certains de ces points pouvant se répéter.

Pour les sauts correspondants de la fonction et de ses dérivées introduisons les notations

$$f^{(l)}(x_j^{(l)} + 0) - f^{(l)}(x_j^{(l)} - 0) = h_j^{(l)}$$

$$(l = 0, 1, \ldots, m;\ j = 1, 2, \ldots, k_l).$$

Définissons la fonction $g(x)$ (*fonction des sauts*) par la formule

$$g(x)=\sum_{s=1}^{s=k_0}\frac{h_s^{(0)}}{\pi}\sigma_0(x-x_s^{(0)})+\sum_{s=1}^{s=k_1}\frac{h_s^{(1)}}{\pi}\sigma_1(x-x_s^{(1)})+\ldots$$
$$\ldots+\sum_{s=1}^{s=k_m}\frac{h_s^{(m)}}{\pi}\sigma_m(x-x_s^{(m)}). \qquad (9)$$

La fonction $g(x)$ jouit des propriétés suivantes:

1) aux points $x_1^{(0)}, x_2^{(0)}, \ldots, x_{k_0}^{(0)}$ la fonction $g(x)$ est discontinue, les sauts en ces points étant égaux aux sauts de la fonction $f(x)$ en points correspondants:

$$g(x_j^{(0)}+0)-g(x_j^{(0)}-0)=\frac{h_j^{(0)}}{\pi}[\sigma_0(x_j-x_j+0)-$$
$$-\sigma_0(x_j-x_j-0)]=\frac{h_j^{(0)}}{\pi}\pi=h_j^{(0)}\ ;$$

2) la dérivée $g^{(l)}(x)$ $(l=1, 2, \ldots, m)$ est discontinue aux points $x_1^{(l)}, x_2^{(l)}, \ldots, x_{k_l}^{(l)}$; de plus

$$g^{(l)}(x_j^{(l)}+0)-g^{(l)}(x_j^{(l)}-0)=$$
$$=\frac{h_j^{(l)}}{\pi}[\sigma_l(x_j^{(l)}-x_j^{(l)}+0)-\sigma_l(x_j^{(l)}-x_j^{(l)}-0)]=\frac{h_j^{(l)}}{\pi}\pi=h_j^{(l)},$$

c'est-à-dire

$$g^{(l)}(x_j+0)-g^{(l)}(x_j-0)=f^{(l)}(x_j+0)-f^{(l)}(x_j-0)\ ;$$

3) pour $x\neq x_j^{(l)}$ la fonction $g(x)$ possède des dérivées continues de tous ordres.

Soit

$$\varphi(x)=f(x)-g(x). \qquad (10)$$

En vertu de la première et la deuxième propriété on a:

$$\varphi^{(l)}(x_j^{(l)}+0)-\varphi^{(l)}(x_j^{(l)}-0)=0 \quad (l=0, 1, 2, \ldots, m),$$

ou

$$\varphi(x)\in\widetilde{C}^{(m)}[-\pi, \pi].$$

Ainsi, pour développer la fonction $f(x)$ on peut faire appel à la série de Fourier (2) à convergence rapide. Remarquons que les déve-

loppements

$$\sigma_0(x-x_s^{(0)}) = \sum_{n=1}^{\infty} \frac{\sin n(x-x_s^{(0)})}{n};$$

$$\sigma_1(x-x_s^{(1)}) = -\sum_{n=1}^{\infty} \frac{\cos n(x-x_s^{(1)})}{n^2};$$

$$\sigma_2(x-x_s^{(2)}) = -\sum_{n=1}^{\infty} \frac{\sin n(x-x_s^{(2)})}{n^3};$$

. . . .

permettent de développer aisément la fonction $g(x)$ en série de Fourier. On a finalement que la série de Fourier de la fonction $f(x)$ est composée: a) de la partie lentement convergente dont la somme est évidemment la fonction $g(x)$ et b) du reste à convergence rapide qui est une série de Fourier de la fonction $\varphi(x) \in \widetilde{C}^{(m)}[-\pi, \pi]$.

R e m a r q u e. Si aux extrémités du segment $[-\pi, \pi]$ les valeurs limites de la fonction $f(x)$ ou de ses dérivées $f'(x), \ldots$ $\ldots, f^{(k)}(x)$ $(k \leqslant m)$ ne coïncident pas, c'est-à-dire si

$$f^{(l)}(-\pi+0) \neq f^{(l)}(\pi-0) \quad (l=0, 1, 2, \ldots, k),$$

les points $x = -\pi$ et $x = \pi$ doivent être considérés comme points de discontinuité de la fonction $f(x)$ ou respectivement des dérivées $f^{(l)}(x)$.

En supposant que la fonction $f(x)$ soit prolongée périodiquement hors du segment $[-\pi, \pi]$ de période 2π on obtient que le saut des dérivées en $x = -\pi$ et $x = \pi$ est le même et égal à

$$h^{(l)} = f^{(l)}(-\pi+0) - f^{(l)}(\pi-0).$$

Par suite de la périodicité de la fonction $\sigma_l(x)$, on a:

$$\sigma_l(x+\pi) = \sigma_l(x-\pi),$$

la fonction $\sigma_l^{(l)}(x+\pi)$ sur le segment $[-\pi, \pi]$ admettant deux points de discontinuité ($x = -\pi$ et $x = \pi$) de même saut égal à π. C'est pourquoi il faut inclure dans la formule (9) un seul point extrême, par exemple $x = -\pi$. En effet, d'après la formule (9), le saut de la dérivée $g^{(l)}(x)$ au point $x = -\pi$ est égal à

$$g^{(l)}(-\pi+0) - g^{(l)}(-\pi-0) = \frac{h^{(l)}}{\pi}[\sigma^{(l)}(+0) - \sigma^{(l)}(-0)] = h^{(l)}.$$

La périodicité de $g^{(l)}(x)$ fait que le saut de cette dérivée est également le même pour $x = \pi$. Donc, en formant la différence

$$f(x) - g(x) = \varphi(x),$$

où on ne tient compte que du point $x = -\pi$, on supprime la discontinuité de la l-ième dérivée de la fonction $\varphi(x)$ en $x = -\pi$, de même qu'en $x = \pi$.

E x e m p l e. Améliorer par la méthode de Krylov la convergence de la série de Fourier de la fonction (fig. 48a)

$$f(x)=\begin{cases} x^2+1 & \text{avec} \quad -\pi<x<0, \\ x^2 & \text{avec} \quad 0<x<\pi. \end{cases}$$

S o l u t i o n. D'après la remarque, la fonction $f(x)$ a sur le segment $[-\pi, \pi]$ pour points de discontinuité $x_1 = -\pi$; $x_2 = 0$;

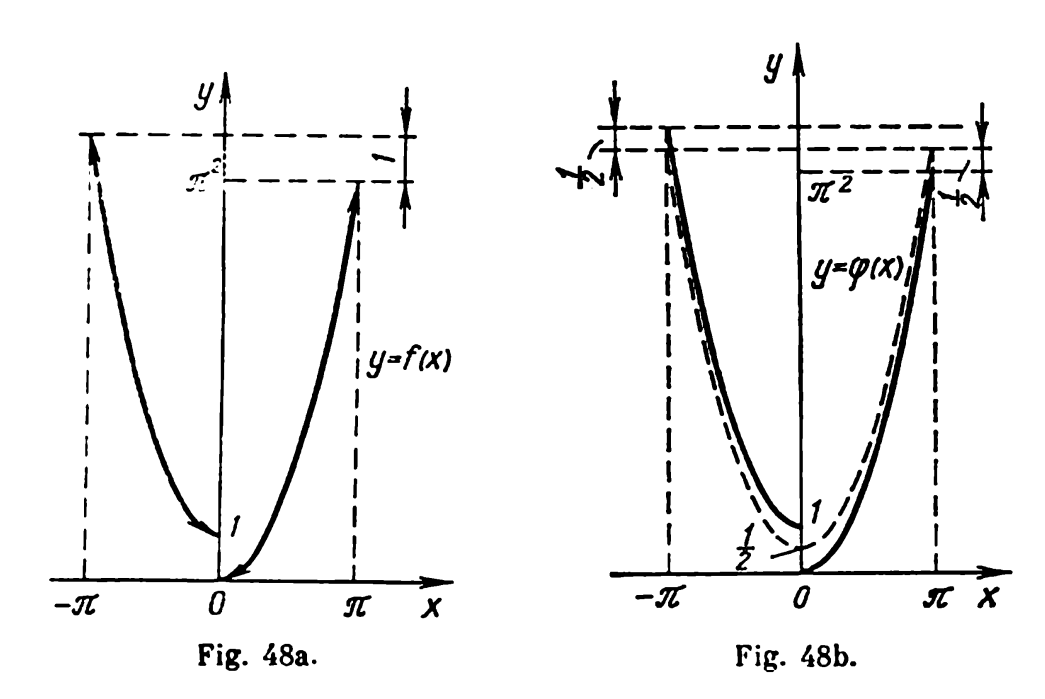


Fig. 48a. Fig. 48b.

$x_3 = \pi$. La calcul des coefficients de Fourier donne

$$a_0 = 1+\frac{2\pi^2}{3}; \quad a_n = \frac{4}{n^2}(-1)^n; \quad b_n = \begin{cases} -\frac{2}{\pi n} & \text{pour } n \text{ impair}; \\ 0 & \text{pour } n \text{ pair}. \end{cases}$$

Par suite, la série de Fourier de la fonction $f(x)$ s'écrit

$$f(x) = \frac{1}{2}+\frac{\pi^2}{3}+4\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos nx - \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{1}{2k+1}\sin(2k+1)x. \quad (11)$$

La convergence de la série (11) est mauvaise du fait que les coefficients $b_n = O\left(\frac{1}{n}\right)$ décroissent lentement. Extrayons de la fonction $f(x)$ la fonction des sauts $g(x)$ de façon que $\varphi(x) = [f(x) - g(x)] \in \widetilde{C}^{(m)}[-\pi, \pi]$.

Calculons les sauts $h_j^{(0)}$ aux points x_j $(j = 1, 2, 3)$:

$$h_1^{(0)} = f(-\pi+0) - f(\pi-0) = (\pi^2+1) - \pi^2 = 1;$$
$$h_2^{(0)} = f(+0) - f(-0) = 0 - 1 = -1;$$
$$h_3^{(0)} = h_1^{(0)} = 1.$$

En vertu de la formule (9) et en tenant compte de la remarque, on obtient

$$g(x) = \frac{1}{\pi}\cdot\sigma_0(x+\pi) - \frac{1}{\pi}\cdot\sigma_0(x)$$

ou

$$g(x) = \frac{1}{\pi}\ \frac{\pi-(x+\pi)}{2} + \frac{1}{\pi}\ \frac{\pi+x}{2} = \frac{1}{2}$$

avec $-\pi < x < 0$ et

$$g(x) = \frac{1}{\pi}\ \frac{\pi-(x+\pi)}{2} - \frac{1}{\pi}\cdot\frac{\pi-x}{2} = -\frac{1}{2}$$

avec $0 < x < \pi$.

En retranchant de la fonction $f(x)$ la fonction des sauts $g(x)$, on obtient la fonction

$$\varphi(x) = x^2 + \frac{1}{2},$$

continue sur le segment $[-\pi, \pi]$ (fig. 48b). Etant donné que

$$\sigma_0(x) = \sum_{n=1}^{\infty} \frac{\sin nx}{n}$$

et que

$$\sigma_0(x+\pi) = \sum_{n=1}^{\infty} \frac{\sin n(x+\pi)}{n} = \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin nx,$$

il vient

$$g(x) = \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin nx - \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin nx =$$
$$= \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n - 1}{n} \sin nx = -\frac{2}{\pi} \sum_{k=0}^{\infty} \frac{\sin(2k+1)x}{2k+1}.$$

Donc

$$f(x) = g(x) + \frac{1}{2} + \frac{\pi^2}{3} + 4 \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos nx,$$

l'ordre de décroissance des coefficients de la série de Fourier transformée étant $O\left(\frac{1}{n^2}\right)$.

Remarquons que si l'on extrait de la fonction $f(x)$ la fonction des sauts $g(x)$ en prenant en considération les discontinuités de la dérivée, le reste sera identiquement nul, on obtient donc la somme exacte de la série (10).

R e m a r q u e. La méthode de Krylov est également applicable aux séries de Fourier de période $T = 2l$. En effet, soit la fonction $f(x)$ définie sur le domaine essentiel $a - l < x < a + l$. Après avoir effectué la transformation linéaire

$$x = a + \frac{l}{\pi} t,$$

on obtient la fonction $F(t) = f\left(a + \frac{l}{\pi} t\right)$ 2π-périodique définie dans le domaine normalisé $-\pi < t < \pi$.

§ 5. Sommation approchée des séries trigonométriques

Soit la série trigonométrique convergente

$$\sum_{n=0}^{\infty} (a_n \cos nx + b_n \sin nx) = S(x), \tag{1}$$

dont la somme $S(x)$ est inconnue. Il faut calculer la valeur approchée de cette somme avec une précision donnée à l'avance.

Il est évident que plus vite les coefficients a_n et b_n de la série (1) tendent vers zéro, moins il faut prendre de termes de la série pour assurer la précision imposée. Pour cette raison avant de commencer le calcul, il convient d'améliorer la convergence de la série. A cette fin on recourt en général à l'artifice suivant: on extrait de la série donnée une certaine série trigonométrique dont la somme $g(x)$ est connue pour que la série restante

$$\sum_{n=0}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx) \tag{2}$$

ait une convergence plus rapide que la série initiale.

Si

$$g(x) = \sum_{n=0}^{\infty} (\overline{a}_n \cos nx + \overline{b}_n \sin nx),$$

il vient

$$S(x) = g(x) + \sum_{n=0}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx), \tag{3}$$

où

$$\alpha_n = a_n - a_n \quad (n = 0, 1, 2, \ldots).$$

Dans les cas les plus simples, pour construire les fonctions $g(x)$ on peut utiliser les développements décrits dans ce qui précède:

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n} = \sigma_0(x) = \frac{\pi - x}{2} \quad (0 < x < 2\pi);$$

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^2} = -\sigma_1(x) = \frac{(\pi - x)^2}{4} - \frac{\pi^2}{12} \quad (0 \leqslant x \leqslant 2\pi);$$

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n^3} = -\sigma_2(x) = \frac{2\pi^2 x - 3\pi x^2 + x^3}{12} \quad (0 \leqslant x \leqslant 2\pi);$$

. . . .

Parfois il est utile également de faire appel aux développements [7]

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n} = -\ln\left(2 \sin \frac{x}{2}\right) \quad (0 < x < 2\pi);$$

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n^2} = -\int_0^x \ln\left(2 \sin \frac{x}{2}\right) dx \quad (0 \leqslant x \leqslant 2\pi);$$

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^3} = \int_0^x dx \int_0^x \ln\left(2 \sin \frac{x}{2}\right) dx + \sum_{n=1}^{\infty} \frac{1}{n^3} \quad (0 \leqslant x \leqslant 2\pi),$$

où $\sum\limits_{n=1}^{\infty} \frac{1}{n^3} = 1{,}202056903\ldots$

E x e m p l e. Trouver la somme de la série

$$S(x) = \sum_{n=1}^{\infty} \frac{n}{n^2+1} \sin nx$$

à 0,001 près.

S o l u t i o n. L'ordre de décroissance des coefficients de la série $b_n = \frac{n}{n^2+1}$ est $O\left(\frac{1}{n}\right)$ puisque $\lim\limits_{n\to\infty}\left(b_n ; \frac{1}{n}\right) = 1$. Améliorons la convergence de la série proposée. Il est clair que

$$\frac{n}{n^2+1} = \frac{n}{n^2}\left(\frac{1}{1+\frac{1}{n^2}}\right) = \frac{1}{n}\left(1 - \frac{1}{n^2} + \frac{1}{n^4} - \ldots\right) = \frac{1}{n} - \frac{1}{n^3} + \gamma_n,$$

où

$$\gamma_n = \frac{n}{n^2+1} - \frac{1}{n} + \frac{1}{n^3} = \frac{1}{n^3(n^2+1)}.$$

Alors,

$$\sum_{n=1}^{\infty} \frac{n}{n^2+1} \sin nx = \sum_{n=1}^{\infty} \frac{\sin nx}{n} - \sum_{n=1}^{\infty} \frac{\sin nx}{n^3} + \sum_{n=1}^{\infty} \gamma_n \sin nx.$$

Mais

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n} = \sigma_0(x) \quad \text{et} \quad \sum_{n=1}^{\infty} \frac{\sin nx}{n^3} = -\sigma_2(x).$$

Donc

$$S(x) = \sigma_0(x) + \sigma_2(x) + \sum_{n=1}^{\infty} \gamma_n \sin nx,$$

où $\gamma_n = \frac{1}{n^3(n^2+1)} = O\left(\frac{1}{n^5}\right)$.

Soit N le nombre de termes de la série $\sum_{n=1}^{\infty} \gamma_n \sin nx$ qu'il faut prendre pour que le reste R_N vérifie l'inégalité

$$|R_N| = \left|\sum_{N+1}^{\infty} \gamma_n \sin nx\right| < 0{,}001.$$

Trouvons le nombre N. On a:

$$\left|\sum_{N+1}^{\infty} \frac{1}{n^3(n^2+1)} \sin nx\right| < \sum_{N+1}^{\infty} \frac{1}{n^5} < \int_{N}^{\infty} \frac{dx}{x^5} = \frac{1}{4N^4}.$$

En résolvant l'inégalité $\frac{1}{4N^4} < 0{,}001$ on voit qu'il suffit de prendre $N = 5$. Donc, on a avec la précision imposée:

$$S(x) = \frac{\pi - x}{2} - \frac{2\pi^2 x - 3\pi x^2 + x^3}{12} + \sum_{n=1}^{5} \frac{\sin nx}{n^3(n^2+1)} \quad (0 < x < \pi).$$

BIBLIOGRAPHIE

1. *G. Fichtengoltz.* Principes d'analyse mathématique, t. II. Gostekhizdat, 1956, chapitres XV et XXIV.
2. *A. Markov.* Calcul des différences finies, 2e éd. Matésis, 1910, chapitre II.
3. *G. Salékhov.* Calcul des séries. Gostekhizdat, 1955, chapitres I et III.
4. *I. Bésikovitch.* Calcul des différences finies. Université d'Etat de Léningrad, 1939, chapitre IX.
5. *A. Guelfond.* Calcul des différences finies. Dunod, 1962, chapitre IV.
6. *Ch.-J. de La Vallée Poussin.* Cours d'analyse infinitésimale, t. II. New York, Dover publications, 1946.
7. *G. Tolstov.* Séries de Fourier. Gostekhizdat, 1951, chapitres I à V.
8. *A. Krylov.* Conférences sur les calculs approchés, 6e éd. Gostekhizdat, 1954, chapitre V.
9. *L. Kantorovitch, V. Krylov.* Méthodes approchées de l'analyse supérieure, 3e éd. Gostekhizdat, 1949, chapitre I.

CHAPITRE VII

ALGÈBRE DES MATRICES

§ 1. Généralités

Un ensemble de mn nombres (réels ou complexes) rangés dans un tableau rectangulaire de m lignes et n colonnes

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{12} & a_{22} & a_{23} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mn} \end{bmatrix} \tag{1}$$

s'appelle *matrice* (numérique).

Les nombres a_{ij} ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$) qui composent la matrice donnée se nomment ses *éléments*. Le premier indice i désigne ici le numéro de la ligne de l'élément, et le deuxième j le numéro de la colonne.

La matrice (1) s'écrit souvent sous une forme condensée

$$A = [a_{ij}] \quad (i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n)$$

ou

$$A = [a_{ij}]_{m \cdot n}.$$

On dit alors que la matrice est d'ordre $m \times n$.

Si $m = n$ la matrice s'appelle *carrée d'ordre n*. Mais si $m \neq n$, on dit que la matrice (1) est *rectangulaire*. En particulier, lorsqu'elle est d'ordre $1 \times n$ on lui donne le nom de *vecteur ligne*, et de *vecteur colonne* si elle est d'ordre $m \times 1$. Un nombre (scalaire) peut être considéré comme une matrice 1×1. Une matrice carrée

$$A=\begin{bmatrix} \alpha_1 & 0 & 0 & \ldots & 0 \\ 0 & \alpha_2 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & & \alpha_n \end{bmatrix} \tag{2}$$

s'appelle matrice *diagonale* et sa notation abrégée est la suivante: $[\alpha_1, \alpha_2, \ldots, \alpha_n]$.

Si $\alpha_i = 1$ $(i = 1, 2, \ldots, n)$, la matrice (2) s'appelle *matrice unité* et on la désigne généralement par le symbole E

$$E = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

En introduisant le *symbole de Kronecker*

$$\delta_{ij} = \begin{cases} 0, & \text{si } i \neq j; \\ 1, & \text{si } i = j, \end{cases}$$

on peut écrire

$$E = [\delta_{ij}].$$

Une matrice dont tous les éléments sont nuls est dite *nulle*; on la désigne par 0. Si l'on veut renseigner encore sur le nombre de ses lignes et de ses colonnes, on écrit 0_{mn}.

A la matrice carrée $A = [a_{ij}]_{n \cdot n}$ est liée la notion du *déterminant*

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}.$$

Il ne faut pas identifier ces deux notions: une matrice est un e n s e m b l e o r d o n n é de nombres mis sous la forme d'un tableau rectangulaire, alors que son déterminant $\det A$ est un nombre, défini par les règles bien connues, et notamment

$$\det A = \sum_{(\alpha_1, \alpha_2, \ldots, \alpha_n)} (-1)^{\varkappa} a_{1\alpha_1} a_{2\alpha_2} \ldots a_{n\alpha_n}, \tag{3}$$

où la somme s'étend à toutes les permutations possibles $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ des éléments $1, 2, \ldots, n$ et, par suite, compte $n!$ termes, de plus $\varkappa = 0$ si la permutation est paire et $\varkappa = 1$ si elle est impaire.

§ 2. Opérations sur les matrices

A. Egalité des matrices

Deux matrices $A = [a_{ij}]$ et $B = [b_{ij}]$ sont considérées comme égales: $A = B$ si elles sont de même type, c'est-à-dire si elles ont le même nombre de lignes et de colonnes et si leurs éléments respectifs sont égaux:

$$a_{ij} = b_{ij}.$$

B. Somme et différence

On appelle *somme* de deux matrices $A = [a_{ij}]$ et $B = [b_{ij}]$ de même ordre une matrice $C = [c_{ij}]$ de même ordre également dont les éléments c_{ij} sont égaux aux sommes des éléments respectifs a_{ij} et b_{ij} des matrices A et B: $c_{ij} = a_{ij} + b_{ij}$. Ainsi

$$A + B = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2n}+b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mn}+b_{mn} \end{bmatrix}.$$

La définition de la somme des matrices entraîne immédiatement les propriétés suivantes:

1) $A + (B + C) = (A + B) + C$;

2) $A + B = B + A$;

3) $A + 0 = A$.

La *différence des matrices* se définit d'une façon analogue

$$A - B = \begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \dots & a_{1n}-b_{1n} \\ a_{21}-b_{21} & a_{22}-b_{22} & \dots & a_{2n}-b_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1}-b_{m1} & a_{m2}-b_{m2} & \dots & a_{mn}-b_{mn} \end{bmatrix}.$$

C. Produit d'une matrice par un nombre

Le *produit d'une matrice* $A = [a_{ij}]$ par un nombre α (ou le produit d'un nombre α par une matrice A) est une *matrice* dont les éléments s'obtiennent par multiplication de tous les éléments de la matrice A par le nombre α, soit

$$A\alpha = \alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha a_{m1} & \alpha a_{m2} & \dots & \alpha a_{mn} \end{bmatrix}.$$

La définition du produit d'un nombre par une matrice rend immédiates les propriétés suivantes:

1) $1A = A$;

2) $0A = 0$;

3) $\alpha(\beta A) = (\alpha\beta)A$;

4) $(\alpha + \beta)A = \alpha A + \beta A$;

5) $\alpha(A + B) = \alpha A + \alpha B$

(ici A et B sont des matrices, α et β des nombres).

15—01072

Remarquons que si la matrice A est une matrice carrée d'ordre n,

$$\det \alpha A = \alpha^n \det A.$$

La matrice

$$-A = (-1)\, A$$

se nomme matrice *opposée*. On voit sans peine que si les matrices A et B sont de même ordre, il vient

$$A - B = A + (-B).$$

D. Multiplication des matrices

Soient

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

et

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1q} \\ b_{21} & b_{22} & \dots & b_{2q} \\ \dots & \dots & \dots & \dots \\ b_{p1} & b_{p2} & \dots & b_{pq} \end{bmatrix}$$

deux matrices respectivement $m \times n$ et $p \times q$. Si le nombre de colonnes de la matrice A est égal au nombre de lignes de la matrice B, c'est-à-dire si

$$n = p, \qquad (1)$$

on définit pour ces matrices la matrice C $m \times q$ dite *produit* des matrices initiales:

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1q} \\ c_{21} & c_{22} & \dots & c_{2q} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mq} \end{bmatrix},$$

où

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{in}b_{nj}$$
$$(i = 1, 2, \dots, m;\ j = 1, 2, \dots, q).$$

Cette définition entraîne la règle suivante de multiplication des matrices: *l'élément de la i-ème ligne et de la j-ième colonne du produit de deux matrices s'obtient en multipliant les éléments de la i-ème ligne de la première matrice par les éléments respectifs de la j-ième colonne de la deuxième matrice et en additionnant les produits obtenus.*

Le produit AB a un sens si et seulement si les lignes de la matrice A comptent autant d'éléments que les colonnes de la matrice B. En particulier, on ne peut multiplier que les matrices carrées de même ordre.

Exemple 1.

$$A=\begin{bmatrix}3 & 2 & 8 & 1\\ 1 & -4 & 0 & 3\end{bmatrix};$$

$$B=\begin{bmatrix}2 & -1\\ 1 & -3\\ 0 & 1\\ 3 & 1\end{bmatrix}.$$

$$AB=$$

$$=\begin{bmatrix}3\cdot 2+2\cdot 1+8\cdot 0+1\cdot 3 & 3\cdot(-1)+2\cdot(-3)+8\cdot 1+1\cdot 1\\ 1\cdot 2+(-4)\cdot 1+0\cdot 0+3\cdot 3 & 1\cdot(-1)+(-4)\cdot(-3)+0\cdot 1+3\cdot 1\end{bmatrix}=$$

$$=\begin{bmatrix}11 & 0\\ 7 & 14\end{bmatrix}.$$

Exemple 2.

$$\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{bmatrix}\cdot\begin{bmatrix}1\\ 2\\ 3\end{bmatrix}=\begin{bmatrix}1\cdot 1+2\cdot 2+3\cdot 3\\ 4\cdot 1+5\cdot 2+6\cdot 3\\ 7\cdot 1+8\cdot 2+9\cdot 3\end{bmatrix}=\begin{bmatrix}14\\ 32\\ 50\end{bmatrix}.$$

Un produit matriciel jouit des propriétés suivantes:

1) $A\,(BC) = (AB)\,C$; 3) $(A + B)\,C = AC + BC$;

2) $\alpha\,(AB) = (\alpha A)\,B$; 4) $C\,(A + B) = CA + CB$

(A, B, C sont des matrices, α est un nombre).

Les égalités 1) à 4) sont entendues dans le sens que si l'un de leurs membres existe, l'autre membre existe également et les deux membres sont égaux entre eux.

Le produit de deux matrices n'est pas commutatif, c'est-à-dire, en général, $AB \neq BA$. Pour s'en assurer il suffit d'examiner les exemples qui suivent.

Exemple 3.

$$A=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix};\quad B=\begin{bmatrix}5 & 6\\ 7 & 8\end{bmatrix},$$

et donc

$$AB=\begin{bmatrix}19 & 22\\ 43 & 50\end{bmatrix};\quad BA=\begin{bmatrix}23 & 34\\ 31 & 46\end{bmatrix}.$$

c'est-à-dire ici $AB \neq BA$.

Bien plus, il se peut que le produit de deux matrices prises dans un certain ordre ait un sens, alors que le produit de ces mêmes matrices prises dans l'ordre inverse n'en a aucun.

Il en est ainsi, par exemple, lorsque

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}; \quad B=\begin{bmatrix}3 & 2 & 1\\ 2 & 1 & 3\\ 4 & 3 & 0\end{bmatrix},$$

du fait que

$$AB=\begin{bmatrix}19 & 13 & 7\\ 46 & 31 & 19\end{bmatrix}, \text{ alors que } BA \text{ n'existe pas}$$

Dans les cas particuliers où $AB = BA$, les matrices A et B sont dites *commutables*. Ainsi on voit aisément que la matrice unité E est commutable avec n'importe quelle matrice carrée A de même ordre, et en outre

$$AE = EA = A.$$

Ainsi, dans la multiplication, la matrice unité E joue le *rôle de l'unité*.

Si A et B sont des matrices du même ordre,

$$\det (AB) = \det (BA) = \det A \cdot \det B.$$

Cette formule se déduit de la règle de multiplication des déterminants.

Pour les matrices de l'exemple 3, on a notamment

$$\begin{vmatrix}19 & 22\\ 43 & 50\end{vmatrix} = \begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix}\begin{vmatrix}5 & 6\\ 7 & 8\end{vmatrix}$$

et

$$\begin{vmatrix}23 & 34\\ 31 & 46\end{vmatrix} = \begin{vmatrix}1 & 2\\ 3 & 4\end{vmatrix}\begin{vmatrix}5 & 6\\ 7 & 8\end{vmatrix}.$$

§ 3. Matrice transposée

En remplaçant dans la matrice $m \times n$

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{m1} & a_{m2} & \dots & a_{mn}\end{bmatrix}$$

les lignes par les colonnes respectives, on obtient une matrice dite *transposée* $n \times m$

$$A' = A^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}.$$

En particulier, pour le vecteur ligne

$$a = [a_1 \ a_2 \ \dots \ a_n],$$

la matrice transposée est le vecteur colonne

$$a' = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

Une matrice transposée jouit des propriétés suivantes:

1) une matrice transposée deux fois se confond avec la matrice initiale

$$A'' = (A')' = A;$$

2) la matrice transposée d'une somme est égale à la somme des matrices transposées des termes de l'addition

$$(A + B)' = A' + B';$$

3) la matrice transposée d'un produit est égale au produit des matrices transposées des facteurs pris dans l'ordre inverse

$$(AB)' = B'A'.$$

En effet, l'élément de la i-ème ligne et de la j-ième colonne de la matrice $(AB)'$ est égal à l'élément de la j-ième ligne et de la i-ème colonne de la matrice AB, soit

$$a_{j1}b_{1i} + a_{j2}b_{2i} + \dots + a_{jn}b_{ni}.$$

Cette dernière expression est évidemment la somme des produits des éléments de la i-ème ligne de la matrice B' par les éléments correspondants de la j-ième colonne de la matrice A', c'est-à-dire elle est égale à l'élément généralisé de la matrice $B'A'$.

Si la matrice A est carrée, il est évident que

$$\det A' = \det A.$$

La matrice $A = [a_{ij}]$ s'appelle *symétrique* si elle coïncide avec sa transposée, c'est-à-dire si

$$A' = A. \tag{1}$$

Il résulte de l'égalité (1) que 1) une matrice symétrique est une matrice carrée ($m = n$) et 2) ses éléments, symétriques par rapport à la diagonale principale, sont égaux entre eux

$$a_{ji} = a_{ij}.$$

Le produit

$$C = AA'$$

est, naturellement, une matrice symétrique, puisque

$$C' = (AA')' = (A')'\, A' = AA' = C.$$

Par exemple,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1^2+2^2+3^2 & 1\cdot4+2\cdot5+3\cdot6 \\ 4\cdot1+5\cdot2+6\cdot3 & 4^2+5^2+6^2 \end{bmatrix} = \begin{bmatrix} 14 & 32 \\ 32 & 77 \end{bmatrix}.$$

§ 4. Matrice inverse

Définition 1. On appelle *matrice inverse* par rapport à la matrice donnée une matrice qui étant multipliée à droite ou à gauche par la matrice initiale donne la matrice unité.

Désignons par A^{-1} la matrice inverse de la matrice A. On a alors par définition:

$$AA^{-1} = A^{-1}A = E, \qquad (1)$$

où E est la matrice unité.

La recherche de l'inverse d'une matrice donnée est dite *inversion* de cette matrice.

Définition 2. Une matrice carrée se nomme *régulière* si son déterminant est différent du zéro.

Dans le cas contraire elle est dite *singulière.*

Théorème. *Toute matrice régulière possède une matrice inverse.*

Démonstration. Soit une matrice régulière d'ordre n:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix},$$

où $\det A = \Delta \neq 0$.

Composons pour la matrice A ce qu'on appelle une *matrice adjointe*

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \cdot & \cdot & \cdot & \cdot \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}, \qquad (2)$$

où A_{ij} sont les cofacteurs (mineurs avec leurs signes) des éléments correspondants a_{ij} $(i, j = 1, 2, \ldots, n)$.

Remarquons que les cofacteurs des éléments des lignes s'inscrivent dans les colonnes correspondantes en effectuant ainsi une opération de transposition.

Divisons tous les éléments de la dernière matrice par le déterminant de la matrice A, c'est-à-dire par Δ:

$$A^* = \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \cdots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \cdots & \frac{A_{n2}}{\Delta} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{A_{1n}}{A} & \frac{A_{2n}}{\Delta} & \cdots & \frac{A_{nn}}{\Delta} \end{bmatrix}. \tag{3}$$

Montrons que la matrice A^* est la matrice inverse cherchée: $A^* = A^{-1}$.

On sait que 1) la somme des produits des éléments d'une certaine ligne ou colonne du déterminant par les cofacteurs de ces éléments est égale au déterminant et 2) la somme des produits des éléments d'une certaine ligne ou colonne du déterminant par les cofacteurs des éléments respectifs d'une autre ligne ou colonne correspondante, est nulle, c'est-à-dire

$$\sum_{k=1}^{n} a_{ik}A_{jk} = \delta_{ij}\Delta \tag{4}$$

et

$$\sum_{k=1}^{n} a_{ki}A_{kj} = \delta_{ij}\Delta, \tag{4'}$$

où

$$\delta_{ij} = \begin{cases} 1 & \text{pour } i = j, \\ 0 & \text{pour } i \neq j. \end{cases}$$

En vertu de ces propriétés, la formation du produit AA^* donne

$$AA^* = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \cdots & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \cdots & \frac{A_{n2}}{\Delta} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \cdots & \frac{A_{nn}}{\Delta} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 1 \end{bmatrix} = E. \tag{5}$$

Ainsi, $AA^* = E$.

L'écriture de la formule (5) peut être bien plus compacte si l'on utilise les notations abrégées

$$A = [a_{ij}] \quad \text{et} \quad A^* = \left[\frac{A_{ji}}{\Delta}\right].$$

Si l'on tient compte de la relation (4), on obtient:

$$AA^* = \left[\sum_{k=1}^{n} a_{ik} \frac{A_{jk}}{\Delta}\right] = [\delta_{ij}] = E.$$

D'une façon analogue on peut prouver que $A^*A = E$.

Par conséquent, $A^* = A^{-1}$, c'est-à-dire

$$A^{-1} = \frac{1}{\Delta} [A_{ji}], \tag{6}$$

avec

$$\Delta = \det A.$$

Remarque 1. Pour la matrice donnée A il n'existe qu'une seule matrice inverse A^{-1}. Bien plus même, toute matrice inverse à droite (inverse à gauche) de la matrice A coïncide avec son inverse A^{-1} (si cette dernière existe).

En effet, si

$$AB = E,$$

en prémultipliant cette égalité par A^{-1}, on obtient:

$$A^{-1}AB = A^{-1}E$$

ou

$$B = A^{-1}.$$

D'une façon analogue on montre que si

$$CA = E,$$

alors $C = A^{-1}$.

C'est pourquoi une seule égalité suffit pour vérifier la relation (1).

Remarque 2. Une matrice carrée singulière ne possède pas de matrice inverse. En effet, la matrice A étant une matrice singulière,

$$\det A = 0.$$

L'égalité (1) implique

$$\det A \cdot \det A^{-1} = \det E = 1,$$

soit

$$0 = 1 ?!,$$

ce qui est impossible. La proposition est ainsi démontrée.

Exemple. Trouver l'inverse de la matrice

$$A=\begin{bmatrix} 1 & 2 & 3 \\ -2 & -4 & -5 \\ 3 & 5 & 6 \end{bmatrix}.$$

Solution. Le déterminant étant

$$\Delta=\begin{vmatrix} 1 & 2 & 3 \\ -2 & -4 & -5 \\ 3 & 5 & 6 \end{vmatrix}=\begin{vmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & -1 & -3 \end{vmatrix}=1\neq 0,$$

la matrice A est régulière.

Composons la matrice adjointe

$$\tilde{A}=\begin{vmatrix} 1 & 3 & 2 \\ -3 & -3 & -1 \\ 2 & 1 & 0 \end{vmatrix}.$$

Divisons tous les éléments de la matrice $\tilde{A}$ par $\Delta = 1$ pour obtenir

$$A^{-1}=\begin{bmatrix} 1 & 3 & 2 \\ -3 & -3 & -1 \\ 2 & 1 & 0 \end{bmatrix}.$$

Il est recommandé de vérifier que réellement

$$AA^{-1} = E.$$

Voici certaines propriétés fondamentales d'une matrice inverse.

1. *Le déterminant d'une matrice inverse est égal à la grandeur inverse du déterminant de la matrice initiale.* En effet, soit

$$A^{-1}A = E.$$

Constatant que le déterminant du produit de deux matrices carrées est égal au produit des déterminants de ces matrices, on a :

$$\det A^{-1} \det A = \det E = 1.$$

Par suite,

$$\det A^{-1}=\frac{1}{\det A}.$$

2. *L'inverse du produit de deux matrices carrées est égale au produit des inverses des matrices facteurs pris dans l'ordre opposé,* soit

$$(AB)^{-1} = B^{-1}A^{-1}.$$

En effet,

$$AB\,(B^{-1}A^{-1}) = A\,(BB^{-1})\,A^{-1} = AEA^{-1} = AA^{-1} = E$$

et

$$(B^{-1}A^{-1})\,AB = B^{-1}\,(A^{-1}A)\,B = B^{-1}EB = B^{-1}B = E.$$

Donc $B^{-1}A^{-1}$ est l'inverse de AB.

Plus généralement :

$$(A_1A_2\,\ldots\,A_p)^{-1} = A_p^{-1}A_{p-1}^{-1}\,\ldots\,A_1^{-1}.$$

3. *La transposée de l'inverse d'une matrice est égale à l'inverse de la transposée de la matrice donnée* :

$$(A^{-1})' = (A')^{-1}.$$

En effet, en transposant la relation principale $A^{-1}A = E$, on obtient :

$$(A^{-1}A)' = A'\,(A^{-1})' = E' = E.$$

On en tire en prémultipliant cette dernière égalité par la matrice $(A')^{-1}$

$$(A')^{-1}\,A'\,(A^{-1})' = (A')^{-1}\,E$$

ou

$$(A^{-1})' = (A')^{-1},$$

ce qu'il fallait démontrer.

R e m a r q u e. Une matrice inverse rend plus facile la résolution des équations matricielles

$$AX = B \quad \text{et} \quad YA = B.$$

En effet, si $\det A \neq 0$, il vient :

$$X = A^{-1}B \quad \text{et} \quad Y = BA^{-1}.$$

§ 5. Puissance d'une matrice

Soit A une matrice carrée. Si p est un nombre naturel, on pose :

$$\underbrace{AA\,\ldots\,A}_{p\text{ fois}} = A^p.$$

On convient de plus que $A^0 = E$, où E est une matrice unité. Si la matrice A est régulière, on peut introduire une puissance négative en la définissant par la relation

$$A^{-p} = (A^{-1})^p.$$

Les puissances d'une matrice aux exposants entiers observent les règles usuelles :

1) $A^pA^q = A^{p+q}$;

2) $(A^p)^q = A^{pq}$.

Il est évident qu'il est impossible d'élever à une puissance une matrice rectangulaire non carrée.

Exemple 1. Soit

$$A=\begin{bmatrix}\alpha_1 & 0 & \dots & 0\\ 0 & \alpha_2 & \dots & 0\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & \alpha_n\end{bmatrix}.$$

Il vient

$$A^p=\begin{bmatrix}\alpha_1^p & 0 & \dots & 0\\ 0 & \alpha_2^p & \dots & 0\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & \alpha_n^p\end{bmatrix}.$$

Exemple 2. Trouver

$$\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix}^2.$$

Solution. On a:

$$\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix}^2=\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix}\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix}=\begin{bmatrix}0&0&1&0\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{bmatrix}.$$

Si A et B sont des matrices carrées de même ordre et si en outre $AB = BA$, on a la formule du binôme de Newton

$$(A+B)^p=\sum_{k=0}^{p} C_p^k A^k B^{p-k}.$$

§ 6. Fonctions rationnelles d'une matrice

Soit

$$X=\begin{bmatrix}x_{11} & x_{12} & \dots & x_{1n}\\ x_{21} & x_{22} & \dots & x_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ x_{n1} & x_{n2} & \dots & x_{nn}\end{bmatrix}$$

une matrice carrée arbitraire d'ordre n. Par analogie avec les formules de l'algèbre élémentaire on définit les fonctions polynômes de la

matrice X:

$$P(X) = A_0X^m + A_1X^{m-1} + \ldots + A_mE \quad (\textit{polynôme droit});$$
$$\widetilde{P}(X) = X^mA_0 + X^{m-1}A_1 + \ldots + EA_m \quad (\textit{polynôme gauche}),$$

où $A_\nu (\nu = 0, 1, \ldots, m)$ sont les matrices $m \times n$ ou respectivement $n \times m$ et E la matrice unité d'ordre n.

D'une façon générale, $P(X) \neq \widetilde{P}(X)$.

On peut introduire également des *fonctions rationnelles* de la matrice X en les définissant par les formules

$$R_1(X) = P(X)[Q(X)]^{-1}$$

et

$$R_2(X) = [Q(X)]^{-1}P(X)$$

avec $P(X)$ et $Q(X)$ des polynômes matriciels et $\det[Q(X)] \neq 0$.

E x e m p l e. Soit

$$P(X) = X^2 + \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} X - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

où X est une matrice variable d'ordre deux. Trouver $P\left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right)$.

S o l u t i o n. On a:

$$P\left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}^2 + \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} =$$
$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}.$$

§ 7. Valeur absolue et norme d'une matrice

L'inégalité

$$A \leqslant B \tag{1}$$

entre les matrices $A = [a_{ij}]$ et $B = [b_{ij}]$ de même ordre signifie que

$$a_{ij} \leqslant b_{ij}. \tag{2}$$

Dans ce sens toutes deux matrices ne sont pas comparables entre elles.

Par *valeur absolue* (*module*) d'une matrice $A = [a_{ij}]$ on entend la matrice

$$|A| = [|a_{ij}|],$$

où $|a_{ij}|$ sont les modules des éléments de la matrice A.

Si A et B sont des matrices pour lesquelles les opérations $A+B$ et AB ont un sens, il vient

a) $|A+B| \leqslant |A| + |B|$;

b) $|AB| \leqslant |A| \cdot |B|$;

c) $|\alpha A| = |\alpha| \, |A|$

(α est un nombre).

En particulier, on obtient:

$$|A^p| \leqslant |A|^p$$

(p est un nombre naturel).

Par *norme d'une matrice* $A = [a_{ij}]$ on entend un nombre réel $\|A\|$ qui satisfait aux conditions:

a) $\|A\| \geqslant 0$, de plus $\|A\| = 0$ si et seulement si $A = 0$;

b) $\|\alpha A\| = |\alpha| \, \|A\|$ (α est un nombre) et, en particulier, $\|-A\| = \|A\|$;

c) $\|A+B\| \leqslant \|A\| + \|B\|$;

d) $\|AB\| \leqslant \|A\| \cdot \|B\|$

(A et B sont des matrices pour lesquelles les opérations correspondantes ont un sens). Pour une matrice carrée on a notamment:

$$\|A^p\| \leqslant \|A\|^p,$$

où p est un nombre naturel.

Notons encore une inégalité importante des normes des matrices A et B de même ordre. En appliquant la condition c) on a:

$$\|B\| = \|A + (B-A)\| \leqslant \|A\| + \|B-A\|.$$

D'où

$$\|A-B\| = \|B-A\| \geqslant \|B\| - \|A\|.$$

D'une manière analogue

$$\|A-B\| \geqslant \|A\| - \|B\|.$$

Par conséquent,

$$\|A-B\| \geqslant |\,\|B\| - \|A\|\,|.$$

Appelons une norme *canonique*, si elle vérifie les conditions supplémentaires:

e) si $A = [a_{ij}]$, on a

$$|a_{ij}| \leqslant \|A\|,$$

en outre, pour une matrice scalaire $A = [a_{11}]$ on a $\|A\| = |a_{11}|$;

f) l'inégalité $|A| \leqslant |B|$ (A et B sont des matrices) conduit à l'inégalité

$$\| A \| \leqslant \| B \|.$$

En particulier, $\| A \| = \| \, | A | \, \|$.

Dans ce qui suit, pour la matrice $A = [a_{ij}]$ d'ordre quelconque nous considérerons essentiellement trois normes facilement calculables:

1) $\| A \|_m = \max\limits_i \sum\limits_j | a_{ij} |$ (*m-norme*);

2) $\| A \|_l = \max\limits_j \sum\limits_i | a_{ij} |$ (*l-norme*);

3) $\| A \|_k = \sqrt{\sum\limits_{i, j} | a_{ij} |^2}$ (*k-norme*).

Exemple. Soit

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

On a:

$$\| A \|_m = \max(1+2+3,\ 4+5+6,\ 7+8+9) = \max(6,\ 15,\ 24) = 24;$$

$$\| A \|_l = \max(1+4+7,\ 2+5+8,\ 3+6+9) = \max(12, 15, 18) = 18;$$

$$\| A \|_k = \sqrt{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2} =$$
$$= \sqrt{1+4+9+16+25+36+49+64+81} = \sqrt{285} \approx 16{,}9.$$

En particulier, pour le vecteur

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

ces normes ont les valeurs suivantes:

$$\| x \|_m = \max_i | x_i |;$$

$$\| x \|_l = | x_1 | + | x_2 | + \ldots + | x_n |;$$

$$\| x \|_k \quad | x | = \sqrt{| x_1 |^2 + | x_2 |^2 + \ldots + | x_n |^2}$$

(*ou module du vecteur*). Si les composantes du vecteur sont réelles, on a simplement

$$\| x \|_k = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Vérifions pour les normes $\|A\|_m$, $\|A\|_l$ et $\|A\|_k$ l'observation des conditions a) à d).

On voit immédiatement que les conditions a) et b) sont observées.

Voyons si ces normes satisfont également à la condition c). Soient $A=[a_{ij}]$ et $B=[b_{ij}]$ deux matrices de même ordre. On a:

$$\|A+B\|_m=\max_i\sum_j|a_{ij}+b_{ij}|\leqslant\max_i\{\sum_j|a_{ij}|+\sum_j|b_{ij}|\}\leqslant$$
$$\leqslant\max_i\sum_j|a_{ij}|+\max_i\sum_j|b_{ij}|=\|A\|_m+\|B\|_m.$$

D'une façon analogue,

$$\|A+B\|_l\leqslant\|A\|_l+\|B\|_l.$$

Ensuite

$$\|A+B\|_k=\sqrt{\sum_{i,j}|a_{ij}+b_{ij}|^2}\leqslant$$
$$\leqslant\sqrt{\sum_{i,j}|a_{ij}|^2+\sum_{i,j}|b_{ij}|^2+2\sum_{i,j}|a_{ij}|\,|b_{ij}|}.$$

En appliquant l'inégalité connue de Cauchy *,

$$\sum_{i,j}|a_{ij}|\,|b_{ij}|\leqslant\sqrt{\sum_{i,j}|a_{ij}|^2}\cdot\sqrt{\sum_{i,j}|b_{ij}|^2},$$

---

* Voici la démonstration de *l'inégalité de Cauchy*:

$$\left|\sum_{s=1}^{n}a_sb_s\right|^2\leqslant\sum_{s=1}^{n}|a_s|^2\cdot\sum_{s=1}^{n}|b_s|^2,$$

où a_s et b_s ($s=1, 2, \ldots, n$) sont des nombres complexes arbitraires. Soit λ une variable réelle. Considérons l'inégalité évidente

$$\sum_{s=1}^{n}|a_s\lambda+b_se^{i\varphi_s}|^2\geqslant 0, \qquad (*)$$

où φ_s sont des nombres réels. En désignant par $\bar{a}_s$ et $\bar{b}_s$ les nombres conjugués à a_s et b_s, on a:

$$|a_s\lambda+b_se^{i\varphi_s}|^2=(a_s\lambda+b_se^{i\varphi_s})(\bar{a}_s\lambda+\bar{b}_se^{-i\varphi_s})=$$
$$=a_s\bar{a}_s\lambda^2+(a_s\bar{b}_se^{-i\varphi_s}+\bar{a}_sb_se^{i\varphi_s})\lambda+b_s\bar{b}_s=$$
$$=|a_s|^2\lambda^2+2\operatorname{Re}(a_s\bar{b}_se^{-i\varphi_s})\lambda+|b_s|^2.$$

L'inégalité $(*)$ prend alors la forme

$$\lambda^2\sum_{s=1}^{n}|a_s|^2+2\lambda\sum_{s=1}^{n}\operatorname{Re}(a_s\bar{b}_se^{-i\varphi_s})+\sum_{s=1}^{n}|b_s|^2\geqslant 0.$$

Si l'on pose

$$\varphi_s=\arg(a_s\bar{b}_s),$$

on a

$$\| A+B \|_k \leqslant \sqrt{\sum_{i,j} |a_{ij}|^2} + \sqrt{\sum_{i,j} |b_{ij}|^2} = \| A \|_k + \| B \|_k .$$

Ainsi la condition c) est remplie pour toutes les trois normes.

Vérifions maintenant l'observation de la condition d). Soient la matrice $A = [a_{ij}]$ d'ordre $m' \times n'$ et la matrice $B = [b_{ij}]$ d'ordre $m'' \times n''$. Pour que la multiplication de la première matrice par la deuxième soit possible, il faut que $m'' = n'$. La matrice AB sera d ordre $m' \times n''$.

On a:

$$\| AB \|_m = \max_i \sum_{j=1}^{n''} \left| \sum_{s=1}^{n'} a_{is} a_{sj} \right| \leqslant \max_i \left\{ \sum_{j=1}^{n''} \sum_{s=1}^{n'} |a_{is}| \, |b_{sj}| \right\} =$$

$$= \max_i \left\{ \sum_{s=1}^{n'} |a_{is}| \sum_{j=1}^{n''} |b_{sj}| \right\} \leqslant \max_i \left\{ \sum_{s=1}^{n'} |a_{is}| \cdot \| B \|_m \right\} =$$

$$= \max_i \left\{ \sum_{s=1}^{n'} |a_{is}| \right\} \cdot \| B \|_m = \| A \|_m \cdot \| B \|_m .$$

---

il vient

$$\operatorname{Re} (a_s \bar{b}_s e^{-i\varphi_s}) = \operatorname{Re} \{ | a_s \bar{b}_s | \, e^{i \arg (a_s \bar{b}_s)} \cdot e^{-i \arg (a_s \bar{b}_s)} \} =$$

$$= \operatorname{Re} \{ | a_s \bar{b}_s | \} = | a_s \bar{b}_s | = | a_s b_s |,$$

par suite,

$$\lambda^2 \sum_{s=1}^{n} |a_s|^2 + 2\lambda \sum_{s=1}^{n} |a_s b_s| + \sum_{s=1}^{n} |b_s|^2 \geqslant 0.$$

Le premier membre de la dernière inégalité étant non négatif par suite de l'inégalité initiale ($\ast$) quels que soient λ réels, l'équation quadratique correspondante ne peut posséder de racines réelles distinctes. C'est pourquoi le discriminant de l'équation est tel que

$$\left\{ \sum_{s=1}^{n} |a_s b_s| \right\}^2 - \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2 \leqslant 0,$$

c'est-à-dire

$$\left\{ \sum_{s=1}^{n} |a_s b_s| \right\}^2 \leqslant \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2 .$$

D'où à plus forte raison

$$\left| \sum_{s=1}^{n} a_s b_s \right|^2 \leqslant \left\{ \sum_{s=1}^{n} |a_s b_s| \right\}^2 \leqslant \sum_{s=1}^{n} |a_s|^2 \cdot \sum_{s=1}^{n} |b_s|^2 .$$

Si les nombres a_s et b_s sont réels, on obtient simplement

$$\left(\sum_{s=1}^{n} a_s b_s \right)^2 \leqslant \sum_{s=1}^{n} a_s^2 \cdot \sum_{s=1}^{n} b_s^2 .$$

D'une façon analogue

$$\| AB \|_l = \max_j \sum_{i=1}^{m'} \left| \sum_{s=1}^{n'} a_{is} b_{sj} \right| \leqslant \max_j \left\{ \sum_{i=1}^{m'} \sum_{s=1}^{n'} |a_{is}|\, |b_{sj}| \right\} =$$

$$= \max_j \left\{ \sum_{s=1}^{n'} |b_{sj}| \sum_{i=1}^{m'} |a_{is}| \right\} \leqslant \max_j \left\{ \sum_{s=1}^{n'} |b_{sj}| \cdot \| A \|_l \right\} =$$

$$= \| A \|_l \cdot \max_j \sum_{s=1}^{n'} |b_{sj}| = \| A \|_l \cdot \| B \|_l .$$

Ensuite

$$\| AB \|_k = \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left| \sum_{s=1}^{n'} a_{is} b_{sj} \right|^2} \leqslant \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left\{ \sum_{s=1}^{n'} |a_{is}|\, |b_{sj}| \right\}^2} .$$

En appliquant l'inégalité de Cauchy et en tenant compte du fait que $m'' = n'$, on a

$$\| AB \|_k \leqslant \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n''} \left\{ \sum_{s=1}^{n'} |a_{is}|^2 \cdot \sum_{t=1}^{m''} |b_{tj}|^2 \right\}} =$$

$$= \sqrt{\sum_{i=1}^{m'} \sum_{s=1}^{n'} |a_{is}|^2 \cdot \sum_{t=1}^{m''} \sum_{j=1}^{n''} |b_{tj}|^2} = \sqrt{\| A \|_k^2 \cdot \| B \|_k^2} = \| A \|_k \cdot \| B \|_k .$$

Les normes considérées vérifient donc la condition d).

Montrons que les normes $\| A \|_m$, $\| A \|_l$, $\| A \|_k$ sont canoniques.

Si a_{pq} est l'élément le plus grand en module de la matrice $A = [a_{ij}]$ d'ordre $m' \times n'$, on a évidemment

$$\| a \|_m \geqslant |a_{p1}| + \ldots + |a_{pq}| + \ldots + |a_{pn'}| \geqslant |a_{pq}| ;$$

$$\| A \|_l \geqslant |a_{1q}| + \ldots + |a_{pq}| + \ldots + |a_{m'q}| \geqslant |a_{pq}|$$

et

$$\| A \|_k = \sqrt{\sum_{i=1}^{m'} \sum_{j=1}^{n'} |a_{ij}|^2} \geqslant |a_{pq}| .$$

Ainsi

$$|a_{ij}| \leqslant |a_{pq}| \leqslant \| A \|_s \qquad (s = m, l, k).$$

Par ailleurs, si $A = [a_{11}]$, il vient

$$\| A \|_m = \| A \|_l = \| A \|_k = |a_{11}| .$$

Ensuite, si $|A| \leqslant |B|$, où $A = [a_{ij}]$ et $B = [b_{ij}]$, on a alors $|a_{ij}| \leqslant |b_{ij}|$. De la définition des normes $\| A \|_m$, $\| A \|_l$ et $\| A \|_k$ il apparaît que les inégalités

$$\| A \|_s \leqslant \| B \|_s \qquad (s = m, l, k)$$

ont lieu.

16—01072

On a en outre pour chacune de ces normes

$$\| A \|_s = \| \, | A | \, \|_s \qquad (s = m,\ l,\ k).$$

La condition f) est donc vérifiée elle aussi.

Ainsi, nous avons donc démontré que les normes $\| A \|_m$, $\| A \|_l$ et $\| A \|_k$ sont canoniques.

Notons que si la matrice E est une matrice unité d'ordre n, on a

$$\| E \|_m = \| E \|_l = 1$$

et

$$\| E \|_k = \sqrt{n}.$$

§ 8. Rang d'une matrice

Soit une matrice rectangulaire

$$A = \begin{bmatrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \dots & \dots \\ a_{m1} & a_{m2} \dots a_{mn} \end{bmatrix}.$$

Si dans cette matrice on choisit d'une façon arbitraire k lignes et k colonnes, où $k \leqslant \min(m, n)$, les éléments disposés à l'intersection de ces lignes et de ces colonnes forment une matrice carrée d'ordre k. Le déterminant de cette matrice s'appelle *mineur* d'ordre k de la matrice A.

Définition. *On appelle rang d'une matrice l'ordre maximal de son mineur non nul.* Autrement dit, le rang d'une matrice A est r si

1) au moins un de ses mineurs d'ordre r est non nul;

2) tous les mineurs d'ordre $r + 1$ et d'ordres plus grands sont nuls.

Le rang d'une matrice nulle, c'est-à-dire composée de zéros, est considéré comme nul. La différence entre le plus petit des nombres m et n et le rang de la matrice s'appelle *défaut* d'une matrice. Si le défaut est nul, le rang de la matrice est maximal pour les matrices de l'ordre considéré.

Pour obtenir le rang d'une matrice il est utile d'observer les règles suivantes:

1) passer des mineurs d'ordres inférieurs (à partir des mineurs du premier ordre, c'est-à-dire des éléments de la matrice) aux mineurs d'ordres plus grands;

2) supposons qu'on ait trouvé le mineur D d'ordre r non nul; il reste alors à calculer les mineurs d'ordre $(r + 1)$ qui encadrent le mineur D. Si tous ces mineurs sont nuls, le rang de la matrice est r; mais si au moins l'un d'eux est non nul, il faut reprendre

l'opération pour ce dernier; dans ce cas le rang de la matrice est évidemment toujours bien supérieur à r.

E x e m p l e. Chercher le rang de la matrice

$$\begin{bmatrix} 2 & -4 & 3 & 1 & 0 \\ 1 & -2 & 1 & -4 & 2 \\ 0 & 1 & -1 & 3 & 1 \\ 4 & -7 & 4 & -4 & 5 \end{bmatrix}.$$

S o l u t i o n. Le mineur d'ordre deux en haut à gauche de cette matrice est nul. Toutefois la matrice contient également d'autres mineurs non nuls d'ordre deux, par exemple

$$D = \begin{vmatrix} -4 & 3 \\ -2 & 1 \end{vmatrix} \neq 0;$$

le mineur d'ordre trois qui l'encadre

$$D' = \begin{vmatrix} 2 & -4 & 3 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{vmatrix} = 1$$

et les deux mineurs de quatrième ordre, qui encadrent le mineur D', sont nuls:

$$\begin{vmatrix} 2 & -4 & 3 & 1 \\ 1 & -2 & 1 & -4 \\ 0 & 1 & -1 & 3 \\ 4 & -7 & 4 & -4 \end{vmatrix} = 0; \qquad \begin{vmatrix} 2 & -4 & 3 & 0 \\ 1 & -2 & 1 & 2 \\ 0 & 1 & -1 & 1 \\ 4 & -7 & 4 & 5 \end{vmatrix} = 0.$$

Le rang de la matrice est donc égal à trois et son défaut est $4 - 3 = 1$.

§ 9. Limite d'une matrice

Soit une suite de matrices

$$A_k = [a_{ij}^{(k)}] \qquad (k = 1, 2, \ldots) \tag{1}$$

de même ordre $m \times n$ ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$). Par *limite* de la suite de matrices A_k on entend la matrice

$$A = \lim_{k \to \infty} A_k = [\lim_{k \to \infty} a_{ij}^{(k)}]. \tag{2}$$

La suite de matrices qui possède une limite s'appelle *convergente*.

L e m m e 1. *Pour qu'une suite de matrices A_k ($k = 1, 2, \ldots$) converge vers une matrice A, il faut et il suffit que*

$$\| A - A_k \| \to 0 \quad \textit{quand} \quad k \to \infty, \tag{3}$$

où $\| A \|$ *est une norme canonique quelconque de la matrice A. De plus,*

$$\lim_{k\to\infty} \| A_k \| = \| A \|.$$

En effet, si

$$A_k \to A = [a_{ij}],$$

on a

$$| a_{ij} - a_{ij}^{(k)} | < \varepsilon \quad \text{pour} \quad k > N(\varepsilon).$$

Il en résulte

$$| A - A_k | < \varepsilon I,$$

où I est une matrice $m \times n$ dont tous les éléments sont égaux à l'unité. En vertu des propriétés d'une norme, on a:

$$\| A - A_k \| \leqslant \varepsilon \| I \| \quad \text{pour} \quad k > N(\varepsilon),$$

donc

$$\lim_{k\to\infty} \| A - A_k \| = 0. \tag{4}$$

Inversement, supposons que la condition (3) soit remplie. Alors on a avec $k > N(\varepsilon)$:

$$| a_{ij} - a_{ij}^{(k)} | \leqslant \| A - A_k \| < \varepsilon$$

et par suite

$$\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij},$$

soit

$$\lim_{k\to\infty} A_k = A.$$

En outre, si $A_k \to A$, on a:

$$| \| A \| - \| A_k \| | \leqslant \| A - A_k \| \to 0 \quad \text{pour} \quad k \to \infty.$$

C'est pourquoi

$$\lim_{k\to\infty} \| A_k \| = \| A \|.$$

Corollaire. La suite $A_k \to 0$ quand $k \to \infty$ si et seulement si

$$\lim_{k\to\infty} \| A_k \| = 0,$$

où $\| A_k \|$ est une norme canonique quelconque.

On montre aisément que si les limites

$$\lim_{k\to\infty} A_k = A \quad \text{et} \quad \lim_{k\to\infty} B_k = B,$$

on aura

$$\text{a) } \lim_{k\to\infty}(A_k \pm B_k) = A \pm B,$$

$$\text{b) } \lim_{k\to\infty}(A_k B_k) = AB,$$

$$\text{c) } \lim_{k\to\infty} A_k^{-1} = A^{-1} \ (\det A \neq 0),$$

sous l'hypothèse que les opérations correspondantes aient un sens. En particulier, si C est une matrice constante telle qu'elle rend possibles les produits CA_k et A_kC ($k = 1, 2, \ldots$), alors

$$\lim_{k\to\infty} CA_k = CA$$

et

$$\lim_{k\to\infty} A_kC = AC.$$

Lemme 2. *Pour que la suite des matrices* A_k ($k = 1, 2, \ldots$) *soit convergente, il faut et il suffit qu'on observe le critère généralisé de Cauchy*, et notamment : pour tout $\varepsilon > 0$ il doit exister un nombre $N = N(\varepsilon)$ tel que pour $k > N$, $p > 0$

$$\| A_{k+p} - A_k \| < \varepsilon, \tag{5}$$

avec $\| A_k \|$ une norme canonique quelconque.

En effet, si l'inégalité (5) est vérifiée, tout élément $a_{ij}^{(k)}$ de la matrice A_k satisfait au critère de Cauchy (cf. chapitre III, § 4) et, par conséquent, il existe une limite

$$\lim_{k\to\infty} A_k = [\lim_{k\to\infty} a_{ij}^{(k)}].$$

Inversement, s'il existe

$$A = \lim_{k\to\infty} A_k,$$

le lemme 1 conduit à

$$\| A - A_k \| \to 0 \quad \text{quand} \quad k \to \infty,$$

et donc l'inégalité (5) a lieu.

§ 10. Séries matricielles

En utilisant la notion de limite d'une matrice on peut introduire dans la discussion les *séries matricielles*

$$\sum_{k=1}^{\infty} A_k = \lim_{N\to\infty} \sum_{k=1}^{N} A_k, \tag{1}$$

où A_k sont les matrices de même ordre.

Si la limite (1) existe, la série matricielle s'appelle *convergente* et la matrice obtenue à la limite sera dite *somme* de cette série. Si la limite (1) n'existe pas, la série matricielle se nomme *divergente* et on ne lui affecte aucune somme.

Condition nécessaire de convergence d'une série matricielle

Théorème 1. *Si la série matricielle* (1) *converge, la limite*

$$\lim_{k\to\infty} A_k = 0.$$

Démonstration. Soit

$$S_k = \sum_{j=1}^{k} A_j.$$

Si la série (1) converge, il existe une limite finie

$$S = \lim_{k\to\infty} S_k.$$

On a

$$A_k = S_k - S_{k-1},$$

d'où

$$\lim_{k\to\infty} A_k = \lim_{k\to\infty} S_k - \lim_{k\to\infty} S_{k-1} = S - S = 0.$$

Si la série

$$\sum_{k=1}^{\infty} |A_k| \tag{2}$$

converge, la série matricielle (1) se nomme *absolument convergente*.

Théorème 2. *Une série matricielle absolument convergente est convergente.*

Démonstration. Soit

$$A_k = [a_{ij}^{(k)}] \qquad (k = 1, 2, \ldots).$$

Donc

$$\sum_{k=1}^{\infty} |A_k| = [\sum_{k=1}^{\infty} |a_{ij}^{(k)}|].$$

La série matricielle (2) étant convergente, toute série numérique $\sum_{k=1}^{\infty} |a_{ij}^{(k)}|$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$ est par définition convergente. On en tire en vertu du théorème connu de la théorie des séries que toutes les séries $\sum_{k=1}^{\infty} a_{ij}$ $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ convergent également, c'est-à-dire qu'il existe une limite

$$S = \lim_{N\to\infty} S_N = \lim_{N\to\infty} \sum_{k=1}^{N} A_k$$

et, par suite, la série matricielle (1) converge.

Pour une analyse grossière de la convergence de la série (1) on peut faire appel à la condition suffisante énoncée ci-dessous.

Théorème 3. *Si* $\| A \|$ *est une norme canonique quelconque et la série numérique*

$$\sum_{k=1}^{\infty} \| A_k \| \tag{3}$$

converge, la série matricielle (1) *converge également et sa convergence est absolue.*

Démonstration. Soit

$$A_k = [a_{ij}^{(k)}] \quad (k = 1, 2, \ldots).$$

Considérons les séries numériques

$$\sum_{k=1}^{\infty} a_{ij}^{(k)} \tag{4}$$

$(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$. Puisque

$$| a_{ij}^{(k)} | \leqslant \| A_k \|,$$

toute série (4) converge, et sa convergence est absolue. Donc, par définition, la série matricielle

$$\sum_{k=1}^{\infty} A_k = [\sum_{k=1}^{\infty} a_{ij}^{(k)}]$$

converge également et, qui plus est, sa convergence est absolue.

Dans les applications une grande importance revient aux *séries matricielles* entières: droites

$$\sum_{k=0}^{\infty} A_k X^k \tag{5}$$

et gauches

$$\sum_{k=0}^{\infty} X^k A_k, \tag{5'}$$

où X est une matrice carrée d'ordre n. Dans le premier cas, les A_k sont des matrices $m \times n$ ou des nombres (par exemple, les A_k peuvent être des vecteurs lignes); dans le deuxième, les A_k sont des matrices $n \times m$ ou des nombres (par exemple, les A_k peuvent être des vecteurs colonnes).

Théorème 4. *Si* r *est le rayon de convergence d'une série scalaire entière*

$$\sum_{k=0}^{\infty} \| A_k \| x^k, \tag{6}$$

où $\| A_k \|$ $(k = 0, 1, 2, \ldots)$ *est une norme canonique quelconque, les séries matricielles entières* (5) *et* (5′) *toujours convergent bien avec*

$$\| X \| < r. \tag{7}$$

En particulier, la série matricielle entière

$$\sum_{k=0}^{\infty} a_k X^k$$

aux constantes a_k $(k = 0, 1, 2, \ldots)$ converge avec

$$\| X \| < r,$$

où r est le rayon de convergence de la série entière

$$\sum_{k=0}^{\infty} | a_k | x^k.$$

Démonstration. Etant donné que

$$\| A_k X^k \| \leqslant \| A_k \| \, \| X \|^k,$$

l'observation de l'inégalité (7) entraîne la convergence de la série

$$\sum_{k=0}^{\infty} \| A_k X^k \|.$$

Par conséquent, en vertu du théorème 3, la série entière (5) converge également.

Un raisonnement analogue est aussi valide pour la série (5′).

La deuxième proposition du théorème se déduit du fait que si a_k est un nombre,

$$\| a_k \| = | a_k |.$$

Théorème 5. *Les progressions géométriques*

$$A + AX + AX^2 + \ldots + AX^k + \ldots \tag{8}$$

et

$$A + XA + X^2A + \ldots + X^kA + \ldots, \tag{8′}$$

dans lesquelles X *est une matrice carrée, convergent si*

$$\| X \| < 1. \tag{9}$$

En outre,

$$\sum_{k=0}^{\infty} AX^k = A(E-X)^{-1}$$

et

$$\sum_{k=0}^{\infty} X^kA = (E-X)^{-1}A.$$

En effet, en vertu de théorème 4 et de la condition (9) la progression géométrique (8) converge, c'est-à-dire il existe une matrice

$$S=\sum_{k=0}^{\infty} AX^k.$$

Considérons l'identité

$$A\,(E+X+X^2+\ldots+X^k)\,(E-X)=A\,(E-X^{k+1}). \qquad (10)$$

En passant à la limite dans l'égalité (10) pour $k \to \infty$ et compte tenu du fait qu'en vertu de la condition (9)

$$X^{k+1} \to 0 \quad \text{quand} \quad k \to \infty,$$

on a

$$S\,(E-X)=AE=A. \qquad (11)$$

En particulier, si l'on pose dans l'égalité (11) $A=E$, on obtient

$$S_1\,(E-X)=E,$$

où

$$S_1=\sum_{k=0}^{\infty} X^k.$$

Il en résulte

$$\det S_1 \cdot \det\,(E-X)=\det E=1.$$

On a donc

$$\det\,(E-X)\neq 0$$

et, par conséquent, la matrice $E-X$ est régulière, c'est-à-dire $(E-X)^{-1}$ existe.

En multipliant les deux membres de l'égalité (11) par $(E-X)^{-1}$ on aura finalement

$$S=\sum_{k=0}^{\infty} AX^k=A\,(E-X)^{-1}.$$

D'une façon analogue on montre que

$$\sum_{k=0}^{\infty} X^kA=(E-X)^{-1}A$$

pour

$$\| X \| < 1.$$

C o r o l l a i r e. Si $\| X \| < 1$, il existe une matrice inverse

$$(E-X)^{-1}=\sum_{k=0}^{\infty} X^k.$$

De plus, si $\| E \| = 1$, on a

$$\| (E-X)^{-1} \| \leqslant \sum_{k=0}^{\infty} \| X \|^k=\frac{1}{1-\| X \|}.$$

R e m a r q u e. Si $\|X\| < 1$, l'évaluation de la norme du reste d'une série matricielle (8) ne présente pas de difficulté.

On a

$$R_k \equiv \|A(E-X)^{-1} - A(E+X+X^2+\ldots+X^k)\| \leqslant$$
$$\leqslant \|A\|\;\|X^{k+1}+X^{k+2}+\ldots\| \leqslant \|A\|(\|X\|^{k+1}+$$
$$+\|X\|^{k+2}+\ldots) = \frac{\|A\|\;\|X\|^{k+1}}{1-\|X\|}.$$

D'une manière analogue on a pour la série (8'):

$$R'_k = \|(E-X)^{-1}A - (E+X+X^2+\ldots+X^k)A\| \leqslant \frac{\|A\|\;\|X\|^{k+1}}{1-\|X\|}.$$

Les séries matricielles permettent de définir les *fonctions transcendantes d'une matrice.* On pose, par exemple,

$$e^X = \sum_{n=0}^{\infty} \frac{X^n}{n!}, \tag{12}$$

et on peut démontrer que pour toute matrice carrée X la série (12) converge.

§ 11. Matrices partitionnées

Soit une certaine matrice A. Décomposons-la en matrices d'ordres inférieurs (sous-matrices: *blocs* ou *parties*) à l'aide de barres horizontales ou verticales. Par exemple

$$A = \left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \end{array}\right],$$

où les blocs sont constitués par les matrices

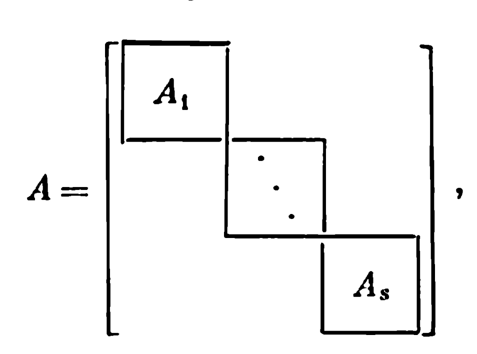
$$P = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}; \quad Q = \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}; \quad R = [a_{31} \;\; a_{32}]; \quad S = [a_{33}].$$

La matrice A peut alors être considérée comme une matrice composée, dont les éléments sont les blocs:

$$A = \begin{bmatrix} P & Q \\ R & S \end{bmatrix}.$$

Une matrice *décomposée en blocs* est dite encore matrice *partitionnée.* Il est clair que la décomposition d'une matrice en blocs peut s'effectuer de façons différentes. Un cas particulier de matrices partitionnées est celui des matrices *quasi diagonales* ou *presque diagonales*

par blocs



dont les blocs A_i ($i = 1, \ldots, s$) sont des matrices carrées, en général, d'ordres différents, alors que hors des blocs figurent des zéros. Constatons que

$$\det A = \det A_1 \ldots \det A_s.$$

Un autre cas important de matrices partitionnées est celui des *matrices encadrées*

$$A_n = \left[\begin{array}{c|c} A_{n-1} & U_n \\ \hline V_n & a_{nn} \end{array}\right],$$

où

$$A_{n-1} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1,\,n-1} \\ a_{21} & a_{22} & & a_{2,\,n-2} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n-1,\,1} & a_{n-1,\,2} & \ldots & a_{n-1,\,n-1} \end{bmatrix}$$

est une matrice d'ordre $n-1$;

$$U_n = \begin{bmatrix} a_{1,\,n} \\ a_{2,\,n} \\ \cdots \\ a_{n-1,\,n} \end{bmatrix}, \text{ une matrice colonne;}$$

$V_n = [a_{n,\,1}\ a_{n,\,2} \ldots a_{n,\,n-1}]$, une matrice ligne et a_{nn} un nombre.

Les matrices partitionnées de même ordre décomposées de la même façon s'appellent par convention *conformes*. La commodité des matrices partitionnées consiste dans le fait que les opérations dont elles sont l'objet se font formellement d'après les mêmes règles que celles relatives aux matrices ordinaires.

A. Addition et soustraction des matrices partitionnées

Si les matrices partitionnées

$$A = \begin{bmatrix} A_{11} & A_{12} \dots & A_{1q} \\ \dots & \dots & \dots \\ A_{p1} & A_{p2} \dots & A_{pq} \end{bmatrix} \quad (1)$$

et

$$B = \begin{bmatrix} B_{11} & B_{12} \dots & B_{1s} \\ \dots & \dots & \dots \\ B_{r1} & B_{r2} \dots & B_{rs} \end{bmatrix} \quad (2)$$

sont conformes, c'est-à-dire si $p = r$; $q = s$ et les blocs A_{ij} et B_{ij} sont de même ordre, il vient

$$A + B = \begin{bmatrix} A_{11}+B_{11} & A_{12}+B_{12} \dots & A_{1q}+B_{1q} \\ \dots & \dots & \dots \\ A_{p1}+B_{p1} & A_{p2}+B_{p2} \dots & A_{pq}+B_{pq} \end{bmatrix}.$$

En effet, pour additionner les matrices A et B il faut additionner leurs éléments respectifs; or il est évident qu'on obtiendra le même résultat en additionnant les blocs respectifs de ces matrices.

La soustraction des matrices partitionnées s'effectue d'une manière analogue.

Si A est une matrice partitionnée (1) et α un nombre, on a

$$\alpha A = \begin{bmatrix} \alpha A_{11} & \alpha A_{12} \dots & \alpha A_{1q} \\ \dots & \dots & \dots \\ \alpha A_{p1} & \alpha A_{p2} \dots & \alpha A_{pq} \end{bmatrix}.$$

B. Multiplication des matrices partitionnées

Soient les matrices partitionnées A et B respectivement à structure (1) et (2); de plus $q = r$.

Supposons que tous les blocs A_{ij} et B_{jk} ($i = 1, 2, \dots, p$; $j = 1, 2, \dots, q$; $k = 1, 2, \dots, s$) sont tels que le nombre de colonnes du bloc A_{ij} est égal au nombre de lignes du bloc B_{jk}. Dans le cas particulier, lorsque tous les blocs A_{ij} et B_{ij} sont carrés et sont de même ordre, cette hypothèse est bien vraie. On peut alors montrer que le produit des matrices A et B est une matrice partitionnée

$$C = \begin{bmatrix} C_{11} & C_{12} \dots & C_{1s} \\ C_{21} & C_{22} \dots & C_{2s} \\ \dots & \dots & \dots \\ C_{p1} & C_{p2} \dots & C_{ps} \end{bmatrix},$$

où $C_{ik} = A_{i1}B_{1k} + A_{i2}B_{2k} + \ldots + A_{iq}B_{qk}$ $(i = 1, 2, \ldots, p;$ $k = 1, 2, \ldots, s)$, c'est-à-dire la multiplication des matrices A et B se fait comme si à la place des blocs il y avait des nombres [2].

Exemple. En multipliant les matrices partitionnées

$$A = \left\| \begin{array}{c|c|c} & \leftarrow 2 \rightarrow & \leftarrow 1 \rightarrow \\ \hline \uparrow\ 2\ \downarrow & P & Q \end{array} \right\|$$

et

$$B = \left\| \begin{array}{c|c|c} & \leftarrow 1 \rightarrow & \leftarrow 2 \rightarrow \\ \hline \uparrow\ 2\ \downarrow & R & S \\ \hline \uparrow\ 1\ \downarrow & T & U \end{array} \right\|,$$

on obtient une matrice de la forme

$$AB = \left\| \begin{array}{c|c|c} & \leftarrow 1 \rightarrow & \leftarrow 2 \rightarrow \\ \hline \uparrow\ 2\ \uparrow & PR + QT & PS + QU \end{array} \right\|.$$

L'addition et la multiplication sont simples surtout lorsqu'il s'agit de matrices quasi diagonales. Si

$$A = \left\| \begin{array}{ccc} \boxed{A_1} & & \\ & \ddots & \\ & & \boxed{A_s} \end{array} \right\|, \qquad B = \left\| \begin{array}{ccc} \boxed{B_1} & & \\ & \ddots & \\ & & \boxed{B_s} \end{array} \right\|$$

et les matrices A_i, B_i $(i = 1, 2, \ldots, s)$ sons de même ordre, on a évidemment

$$A + B = \begin{bmatrix} A_1 + B_1 & & \\ & \ddots & \\ & & A_s + B_s \end{bmatrix}$$

et

$$AB = \begin{bmatrix} A_1B_1 & & \\ & \ddots & \\ & & A_sB_s \end{bmatrix}.$$

§ 12. Inversion des matrices par partition

Supposons qu'il faille trouver pour une matrice numérique régulière A la matrice inverse A^{-1}. Décomposons la matrice A en quatre blocs

$$A = \begin{bmatrix} \alpha_{11}(r, r) & \alpha_{12}(r, s) \\ \alpha_{21}(s, r) & \alpha_{22}(s, s) \end{bmatrix}.$$

Ici entre parenthèses on indique les ordres des blocs correspondants; en outre, $r + s = n$, où n est l'ordre de la matrice A. Cherchons la matrice inverse A^{-1} également sous la forme d'une matrice à quatre blocs

$$A^{-1} = \begin{bmatrix} \beta_{11}(r, r) & \beta_{12}(r, s) \\ \beta_{21}(s, r) & \beta_{22}(s, s) \end{bmatrix}.$$

Etant donné que $A^{-1}A = E$, en multipliant ces matrices, on obtient quatre équations matricielles

$$\left.\begin{aligned} \beta_{11}\alpha_{11} + \beta_{12}\alpha_{21} &= E_r, \\ \beta_{11}\alpha_{12} + \beta_{12}\alpha_{22} &= 0, \\ \beta_{21}\alpha_{11} + \beta_{22}\alpha_{21} &= 0, \\ \beta_{21}\alpha_{12} + \beta_{22}\alpha_{22} &= E_s, \end{aligned}\right\} \qquad (1)$$

où E_r et E_s sont des matrices unités d'ordres correspondants. Après avoir résolu ce système, on détermine les blocs de la matrice A^{-1}. Pour résoudre le système (1) utilisons la méthode d'élimination des inconnues. En multipliant à droite la première équation du système (1) par $\alpha_{11}^{-1}\alpha_{12}$ et en retranchant du produit la deuxième équation de ce système, on obtient:

$$\beta_{12}\left(\alpha_{21}\alpha_{11}^{-1}\alpha_{12}-\alpha_{22}\right)=\alpha_{11}^{-1}\alpha_{12}.$$

On en tire

$$\beta_{12}=-\alpha_{11}^{-1}\alpha_{12}\left(\alpha_{22}-\alpha_{21}\alpha_{11}^{-1}\alpha_{12}\right)^{-1}$$

et

$$\beta_{11}=\alpha_{11}^{-1}-\beta_{12}\alpha_{21}\alpha_{11}^{-1}.$$

D'une façon analogue la troisième et la quatrième équation du système (1) amènent

$$\beta_{22}=\left(\alpha_{22}-\alpha_{21}\alpha_{11}^{-1}\alpha_{12}\right)^{-1}$$

et

$$\beta_{21}=-\beta_{22}\alpha_{21}\alpha_{11}^{-1}.$$

Bien sûr nous supposons ici que les opérations correspondantes aient un sens. Introduisons les matrices

$$\left.\begin{aligned} X&=\alpha_{11}^{-1}\alpha_{12}, & Y&=\alpha_{21}\alpha_{11}^{-1},\\ \theta&=\alpha_{22}-\alpha_{21} & X&=\alpha_{22}-Y\alpha_{12}. \end{aligned}\right\} \qquad (2)$$

L'écriture des blocs β_{ij} $(i, j = 1, 2)$ peut être alors simplifiée:

$$\begin{aligned} \beta_{11}&=\alpha_{11}^{-1}+X\theta^{-1}Y,\\ \beta_{12}&=-X\theta^{-1},\\ \beta_{21}&=-\theta^{-1}Y,\ \beta_{22}=\theta^{-1}. \end{aligned}$$

Les formules (1) définissent les blocs de la matrice A^{-1} sous la condition que α_{11}^{-1} et θ^{-1} existent. Les calculs deviennent plus commodes si on dresse le schéma suivant [4]:

	α_{21}	α_{22}
$X=\alpha_{11}^{-1}\alpha_{12}$	α_{11}^{-1}	α_{12}
θ^{-1}	$Y=\alpha_{21}\alpha_{11}^{-1}$	$\theta=\alpha_{22}-Y\alpha_{12}$

$$A^{-1}=\left[\begin{array}{c|c} \alpha_{11}^{-1}+X\theta^{-1}Y & -X\theta^{-1}\\ \hline -\theta^{-1}Y & \theta^{-1} \end{array}\right].$$

L'application de cette méthode est utile si la matrice α_{11} est facilement inversible.

E x e m p l e 1. Inverser la matrice

$$\begin{bmatrix} 1 & 0 & 3 & -4 \\ 0 & 1 & 5 & 6 \\ -3 & 4 & 0 & 2 \\ -5 & -6 & 2 & 0 \end{bmatrix}.$$

S o l u t i o n. Posons

$$\alpha_{11}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \alpha_{12}=\begin{bmatrix} 3 & -4 \\ 5 & 6 \end{bmatrix};$$

$$\alpha_{21}=\begin{bmatrix} -3 & 4 \\ -5 & -6 \end{bmatrix}; \quad \alpha_{22}=\begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}.$$

En appliquant le schéma donné dans ce qui précède, on a

			—3	4	0	2
			—5	—6	2	0
X	3	—4	1	0	3	—4
	5	6	0	1	5	6
$\theta^{-1}\ \frac{1}{1422}$	16	34	—3	4	—11	—34
	—47	—11	—5	—6	47	16
			Y		0	

Il en résulte

$$X\theta^{-1}=\frac{1}{1422}\begin{bmatrix} 3 & -4 \\ 5 & 6 \end{bmatrix}\begin{bmatrix} 16 & 34 \\ -47 & -11 \end{bmatrix}=\frac{1}{1422}\begin{bmatrix} 236 & 146 \\ -202 & 104 \end{bmatrix},$$

$$\theta^{-1}Y=\frac{1}{1422}\begin{bmatrix} 16 & 34 \\ -47 & -11 \end{bmatrix}\begin{bmatrix} -3 & 4 \\ -5 & -6 \end{bmatrix}=\frac{1}{1422}\begin{bmatrix} -218 & -140 \\ 196 & -122 \end{bmatrix},$$

$$X\theta^{-1}Y=\frac{1}{1422}\begin{bmatrix} 236 & 146 \\ -202 & 104 \end{bmatrix}\begin{bmatrix} -3 & 4 \\ -5 & -6 \end{bmatrix}=\frac{1}{1422}\begin{bmatrix} -1438 & 68 \\ 86 & -1432 \end{bmatrix}.$$

Pour vérifier, on calcule le produit $X\theta^{-1}Y$ suivant deux procédés:

$$X\theta^{-1}Y=(X\theta^{-1})\,Y \quad \text{et} \quad Y\theta^{-1}Y=X\,(\theta^{-1}Y).$$

D'après le schéma général, on a

$$A^{-1} = \frac{1}{1422}\left[\begin{array}{rr|rr} -16 & 68 & -236 & -146 \\ 86 & -10 & 202 & -104 \\ \hline 218 & 140 & 16 & 34 \\ -196 & 122 & -47 & -11 \end{array}\right].$$

Un cas particulier de la méthode exposée est ce qu'on appelle la *méthode d'encadrement*, dont voici le principe. Soit la matrice

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{n1} & \dots & a_{nn} \end{bmatrix}.$$

Composons la suite de matrices

$$S_1 = [a_{11}];$$

$$S_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix};$$

$$S_3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \left[\begin{array}{cc|c} \multicolumn{2}{c|}{S_2} & \begin{array}{c} a_{13} \\ a_{23} \end{array} \\ \hline a_{31} & a_{32} & a_{33} \end{array}\right];$$

$$S_4 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \left[\begin{array}{ccc|c} \multicolumn{3}{c|}{S_3} & \begin{array}{c} a_{14} \\ a_{24} \\ a_{34} \end{array} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right],$$

etc. Chaque matrice suivante s'obtient de la précédente par encadrement. L'inverse de la deuxième de ces matrices S_2^{-1} s'obtient immédiatement

$$S_2^{-1} = \begin{bmatrix} \frac{a_{22}}{\Delta} & -\frac{a_{12}}{\Delta} \\ -\frac{a_{21}}{\Delta} & \frac{a_{11}}{\Delta} \end{bmatrix},$$

où

$$\Delta = a_{11}a_{22} - a_{12}a_{21}.$$

En appliquant à S_3 le schéma de calcul donné ci-dessus on peut obtenir à l'aide de la matrice S_2^{-1} la matrice S_3^{-1}, puis utiliser S_3^{-1} pour obtenir d'une façon analogue S_4^{-1} et enfin $S_n^{-1} = A^{-1}$.

Si l'une des matrices intermédiaires S_i est singulière, la méthode d'encadrement devient inefficace. Pour pallier à cet inconvénient, il faut permuter les lignes de la matrice [5].

E x e m p l e 2. Trouver l'inverse de la matrice

$$A=\begin{bmatrix}1 & 4 & 1 & 3\\ 0 & -1 & 3 & -1\\ 3 & 1 & 0 & 2\\ 1 & -2 & 5 & 1\end{bmatrix}.$$

S o l u t i o n. Ici

$$S_2=\begin{bmatrix}1 & 4\\ 0 & -1\end{bmatrix}=S_2^{-1}.$$

Pour calculer S_3^{-1} on recourt au schéma suivant :

		3	1	0
X	13	1	4	1
	-3	0	-1	3
θ^{-1}	$-\frac{1}{36}$	3	11	-36
		Y		θ

$$X\theta^{-1}Y=\begin{bmatrix}-\frac{13}{12} & -\frac{143}{36}\\ \frac{1}{4} & \frac{11}{12}\end{bmatrix}.$$

Par conséquent,

$$S_3^{-1}=\left[\begin{array}{cc|c}-\frac{1}{12} & \frac{1}{36} & \frac{13}{36}\\ \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12}\\ \hline \frac{1}{12} & \frac{11}{36} & -\frac{1}{36}\end{array}\right].$$

Le calcul de S_4^{-1} se fait d'après le schéma:

$$
\begin{array}{c|c|ccc|c|}
\cline{3-6}
 & & 1 & -2 & 5 & 1 \\
\cline{2-6}
 & \frac{4}{9} & -\frac{1}{12} & \frac{1}{36} & \frac{13}{36} & 3 \\
X & \frac{2}{3} & \frac{1}{4} & -\frac{1}{12} & -\frac{1}{12} & -1 \\
 & -\frac{1}{9} & \frac{1}{12} & \frac{11}{36} & \frac{1}{36} & 2 \\
\cline{2-6}
\theta^{-1} & \frac{9}{22} & -\frac{1}{6} & \frac{31}{18} & \frac{7}{18} & \frac{22}{9} \\
\cline{2-6}
\multicolumn{2}{c}{} & \multicolumn{3}{c}{Y} & \multicolumn{1}{c}{\theta}
\end{array}
$$

$$
X\theta^{-1}Y = \begin{bmatrix} -\frac{1}{33} & \frac{31}{99} & \frac{7}{99} \\ -\frac{1}{22} & \frac{31}{66} & \frac{7}{66} \\ \frac{1}{132} & -\frac{31}{396} & -\frac{7}{396} \end{bmatrix}.
$$

Donc

$$
S_4^{-1} = A^{-1} = \left[\begin{array}{ccc|c} -\frac{5}{44} & \frac{15}{44} & \frac{19}{44} & -\frac{2}{11} \\ \frac{9}{44} & \frac{17}{44} & \frac{1}{44} & -\frac{3}{11} \\ \frac{4}{44} & \frac{10}{44} & -\frac{2}{44} & \frac{1}{22} \\ \hline \frac{3}{44} & -\frac{31}{44} & -\frac{7}{44} & \frac{9}{22} \end{array}\right] =
$$

$$
= \frac{1}{44}\begin{bmatrix} -5 & 15 & 19 & -8 \\ 9 & 17 & 1 & -12 \\ 4 & 10 & -2 & 2 \\ 3 & -31 & -7 & 18 \end{bmatrix}.
$$

§ 13. Matrices triangulaires

Définition. Une matrice carrée s'appelle *triangulaire* si ses éléments au-dessus ou au-dessous de la diagonale principale sont nuls. Par exemple,

$$
T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & t_{nn} \end{bmatrix},
$$

où $t_{ij} = 0$ avec $i > j$, est une matrice triangulaire supérieure. D'une façon analogue

$$T_1 = \begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{21} & t_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ t_{n1} & t_{n2} & & t_{nn} \end{bmatrix},$$

où $t_{ij} = 0$ pour $j > i$, est une matrice triangulaire inférieure.

Une matrice diagonale est un cas particulier d'une matrice supérieure ou inférieure. Le déterminant d'une matrice triangulaire est égal au produit de ses éléments diagonaux, et notamment: si $T = [t_{ij}]$ est une matrice triangulaire, il est clair que $\det T = t_{11}t_{22} \dots t_{nn}$. Aussi une matrice triangulaire n'est-elle régulière que lorsque ses éléments diagonaux ne sont pas tous nuls.

On peut montrer que 1) la somme et le produit des matrices triangulaires de même ordre et de même structure, c'est-à-dire seulement des matrices supérieures ou seulement des matrices inférieures, sont également une matrice triangulaire de même ordre et de même structure; 2) l'inverse d'une matrice triangulaire régulière est également une matrice triangulaire de même ordre et de même structure. Cette dernière circonstance rend facile l'inversion d'une matrice triangulaire.

E x e m p l e 1. Inverser la matrice

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix}.$$

S o l u t i o n. Posons

$$A^{-1} = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix}.$$

Le produit des matrices A et A^{-1} donne:

$$\left.\begin{array}{rr} t_{11} = 1, & t_{11} + 2t_{21} + 3t_{31} = 0, \\ t_{11} + 2t_{21} = 0, & 2t_{22} + 3t_{32} = 0, \\ 2t_{22} = 1, & 3t_{33} = 1. \end{array}\right\}$$

On en tire successivement:

$$t_{11} = 1; \quad t_{21} = -\frac{1}{2}; \quad t_{22} = \frac{1}{2};$$

$$t_{31} = 0; \quad t_{32} = -\frac{1}{3}; \quad t_{33} = \frac{1}{3}.$$

Donc

$$A^{-1}=\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Le théorème qui suit [3] est très important.

Théorème. *Toute matrice carrée*

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix},$$

aux mineurs diagonaux principaux non nuls

$$\Delta_1=a_{11}\neq 0; \quad \Delta_2=\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}\neq 0; \quad \dots; \quad \Delta_n=|A|\neq 0$$

peut être mise sous la forme d'un produit de deux matrices triangulaires de structures différentes (inférieure ou supérieure), cette décomposition étant unique si l'on fixe à l'avance les éléments diagonaux de l'une des matrices triangulaires (si on les pose, par exemple, égaux à un).

Sans démontrer le théorème, bornons-nous seulement à indiquer le moyen d'obtenir les éléments des matrices triangulaires. Soit

$$A = T_1T_2, \tag{1}$$

où

$$T_1 = [b_{ij}], \quad b_{ij} = 0 \quad \text{pour} \quad j > i, \tag{2}$$

est une matrice triangulaire inférieure d'ordre n;

$$T_2 = [c_{ij}], \quad c_{ij} = 0 \quad \text{pour} \quad i > j, \tag{3}$$

est une matrice triangulaire supérieure d'ordre n. D'après la formule (1), le produit de ces matrices donne

$$\sum_{k=1}^{n} b_{ik}c_{kj}=a_{ij} \quad (i, j=1, 2, \dots, n). \tag{4}$$

Les conditions (2) et (3) mettent le système (4) sous la forme

$$\sum_{k=1}^{j} b_{ik}c_{kj}=a_{ij} \quad \text{pour} \quad i\geqslant j \quad (j=1, 2, \dots, n) \tag{4'}$$

et

$$\sum_{k=1}^{i} b_{ik}c_{kj}=a_{ij} \quad \text{pour} \quad i<j \quad (i=1, 2, \dots, n-1). \tag{4''}$$

Par suite de leur structure particulière, les systèmes (4′) et (4″) se résolvent facilement aux éléments diagonaux b_{ii} et c_{ii} près. Pour rendre la solution plus concrète, on peut poser $c_{ii} = 1$ ($i = 1, 2, \ldots, n$).

E x e m p l e 2. Mettre la matrice

$$A = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{bmatrix}$$

sous la forme d'un produit de deux matrices triangulaires T_1 et T_2.

S o l u t i o n. $A = T_1 T_2$. Cherchons T_1 et T_2 sous la forme

$$T_1 = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \quad \text{et} \quad T_2 = \begin{bmatrix} 1 & r_{12} & r_{13} \\ 0 & 1 & r_{23} \\ 0 & 0 & 1 \end{bmatrix}.$$

On a

$$\begin{bmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{11}r_{12} & t_{11}r_{13} \\ t_{21} & t_{21}r_{12} + t_{22} & t_{21}r_{13} + t_{22}r_{23} \\ t_{31} & t_{31}r_{12} + t_{32} & t_{31}r_{13} + t_{32}r_{23} + t_{33} \end{bmatrix};$$

d'où

$$t_{11} = 1; \quad t_{11}r_{12} = -1; \quad t_{11}r_{13} = 2;$$
$$t_{21} = -1; \quad t_{21}r_{12} + t_{22} = 5; \quad t_{21}r_{13} + t_{22}r_{23} = 4;$$
$$t_{31} = 2; \quad t_{31}r_{12} + t_{32} = 4; \quad t_{31}r_{13} + t_{32}r_{23} + t_{33} = 14.$$

La résolution du système amène

$$t_{11} = 1; \quad t_{21} = -1; \quad t_{31} = 2;$$
$$t_{22} = 4; \quad t_{32} = 6; \quad t_{33} = 1;$$
$$r_{12} = -1; \quad r_{13} = 2; \quad r_{23} = \frac{3}{2}.$$

Ainsi

$$T_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 4 & 0 \\ 2 & 6 & 1 \end{bmatrix}$$

et

$$T_2 = \begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & \frac{3}{2} \\ 0 & 0 & 1 \end{bmatrix}.$$

En appliquant la représentation d'une matrice carrée A (det $A \neq \neq 0$) sous la forme d'un produit de deux matrices triangulaires, on peut indiquer encore un procédé de calcul d'une matrice inverse A^{-1}, et notamment, si

$$A = T_1 T_2,$$

il vient

$$A^{-1} = T_2^{-1} T_1^{-1}.$$

Nous avons vu dans ce qui précède que le calcul des inverses des matrices triangulaires est relativement simple.

§ 14. Transformations élémentaires des matrices

Les transformations suivantes s'appellent *transformations élémentaires*:

1) permutation de deux lignes ou de deux colonnes;

2) multiplication de tous les éléments d'une ligne (colonne) par le même nombre non nul;

3) addition aux éléments d'une ligne (colonne) des produits d'éléments correspondants d'une autre ligne (colonne) par un même nombre.

Deux matrices se nomment *équivalentes* si l'une s'obtient de l'autre à la suite d'un nombre fini de transformations élémentaires. Ces matrices ne sont pas en général égales entre elles, mais on peut démontrer qu'elles ont le même rang [6].

On voit aisément que chaque transformation élémentaire d'une matrice carrée A est équivalente au produit de cette dernière par une certaine matrice régulière. En outre, si la transformation porte sur les lignes (colonnes) de la matrice A, le multiplicateur doit être à gauche (à droite) et constituer le résultat d'une application de la transformation élémentaire correspondante à la matrice unité [6]. Par exemple, en permutant dans la matrice

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

la deuxième et la troisième ligne, on obtient une matrice équivalente

$$\breve{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

Cette même matrice $\breve{A}$ s'obtient si dans la matrice unité

$$E=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

on permute la deuxième et la troisième ligne pour multiplier à gauche la matrice ainsi obtenue

$$\breve{E}=\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 1\\ 0 & 1 & 0\end{bmatrix}$$

par la matrice A, c'est-à-dire $\breve{A}=\breve{E}A$.

Les autres transformations élémentaires s'effectuent d'une façon analogue. Remarquons que si dans l'égalité $AA^{-1} = E$ on réalise des transformations identiques des lignes des matrices A et E tant que la matrice A ne se transforme en matrice unité, on aboutira à $\breve{E}AA^{-1} = \breve{E}$, où $\breve{E}$ est la transformée de la matrice unité. Puisque $\breve{E}A = E$, on en tire que $A^{-1} = \breve{E}$, c'est-à-dire que la matrice inverse A^{-1} est la transformée de la matrice unité. Ce principe est à la base du calcul d'une matrice inverse à l'aide de la transformation des lignes [4].

§ 15. Calcul des déterminants

Les transformations élémentaires d'une matrice fournissent le moyen le plus commode pour calculer son déterminant. Soit, par exemple

$$\Delta_n=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix}. \qquad (1)$$

En supposant que $a_{11}\neq 0$, on aura

$$\Delta_n=a_{11}\begin{bmatrix}1 & a_{12} & \ldots & a_{1n}\\ \frac{a_{21}}{a_{11}} & a_{22} & \ldots & a_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{a_{n1}}{a_{11}} & a_{n2} & \ldots & a_{nn}\end{bmatrix}.$$

D'où en retranchant des éléments a_{ij} de la j-ième colonne ($j \geqslant 2$) es éléments respectifs de la première colonne, multipliés par a_{1j},

on obtient

$$\Delta_n = a_{11} \begin{vmatrix} 1 & 0 & \dots & 0 \\ \frac{a_{21}}{a_{11}} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ \frac{a_{n1}}{a_{11}} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix} = a_{11}\Delta_{n-1},$$

où

$$\Delta_{n-1} = \begin{vmatrix} a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ \dots & \dots & \dots & \dots \\ a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix} \tag{2}$$

et

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \quad (i,\ j = 2, 3, \dots, n).$$

Procédons de même pour le déterminant Δ_{n-1}. Si tout élément

$$a_{ii}^{(i-1)} \neq 0 \quad (i = 1, 2, \dots, n),$$

on obtient finalement

$$\Delta_n = a_{11}a_{22}^{(1)} \dots a_{nn}^{(n-1)}. \tag{3}$$

Si l'élément supérieur gauche $a_{k+1,\,k+1}^{(k)}$ d'un déterminant intermédiaire quelconque Δ_{n-k} s'annule, les lignes et les colonnes du déterminant Δ_{n-k} doivent être permutées de façon que l'élément nécessaire soit non nul (ce qui est toujours possible si le déterminant $\Delta \neq 0$). Il faut tenir compte, certes, de la variation du signe du déterminant Δ_{n-k}. On peut établir une règle plus générale. Supposons que le déterminant $\tilde{\Delta}_n = \det[\alpha_{ij}]$ est transformé de façon que $\alpha_{pq} = 1$ (α_{pq} est le « pivot »), c'est-à-dire

$$\tilde{\Delta}_n = \begin{vmatrix} \alpha_{11} & \dots & \alpha_{1q} & \dots & \alpha_{1j} & \dots & \alpha_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_{i1} & \dots & \boxed{\alpha_{iq}} & \dots & \alpha_{ij} & \dots & \alpha_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_{p1} & \dots & 1 & \dots & \boxed{\alpha_{pj}} & \dots & \alpha_{pn} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_{n1} & \dots & \alpha_{nq} & \dots & \alpha_{nj} & \dots & \alpha_{nn} \end{vmatrix}.$$

Il vient alors

$$\tilde{\Delta}_n = (-1)^{p+q}\,\tilde{\Delta}_{n-1},$$

où $\tilde{\Delta}_{n-1} = \det [\alpha_{ij}^{(1)}]$ est le déterminant d'ordre $(n-1)$ qui s'obtient de Δ_n en éliminant la p-ième ligne et la q-ième colonne avec la transformation ultérieure des éléments d'après la formule

$$\alpha_{ij}^{(1)} = \alpha_{ij} - \alpha_{iq}\alpha_{pj},$$

c'est-à-dire *tout élément* $\alpha_{ij}^{(1)}$ *du déterminant* $\tilde{\Delta}_{n-1}$ *est égal à l'élément associé* α_{ij} *du déterminant* $\tilde{\Delta}_n$ *diminué d'un produit de ses « projections »* α_{iq} *et* α_{pj} *sur la colonne et la ligne éliminées du déterminant initial.* Cette proposition est démontrée facilement à partir des propriétés générales des déterminants [7].

Exemple. Calculer

$$\Delta_5 = \begin{vmatrix} 3 & 1 & -1 & 2 & \boxed{1} \\ -2 & 3 & 1 & 4 & 3 \\ 1 & 4 & 2 & 3 & 1 \\ 5 & -2 & -3 & 5 & -1 \\ -1 & 1 & 2 & 3 & 2 \end{vmatrix}.$$

Solution. En prenant comme pivot $a_{15} = 1$, on a

$$\Delta_5 = (-1)^{1+5} \times$$

$$\times \begin{vmatrix} -2-3\cdot 3 & 3-1\cdot 3 & 1-(-1)\cdot 3 & 4-2\cdot 3 \\ 1-3\cdot 1 & 4-1\cdot 1 & 2-(-1)\cdot 1 & 3-2\cdot 1 \\ 5-3\cdot(-1) & -2-1\cdot(-1) & -3-(-1)(-1) & 5-2\cdot(-1) \\ -1-3\cdot 2 & 1-1\cdot 2 & 2-(-1)2 & 3-2\cdot 2 \end{vmatrix} =$$

$$= \begin{vmatrix} -11 & 0 & 4 & -2 \\ -2 & 3 & 3 & \boxed{1} \\ 8 & -1 & -4 & 7 \\ -7 & -1 & 4 & -1 \end{vmatrix}.$$

Ensuite, en prenant pour pivot $a_{24} = 1$ et en appliquant une transformation analogue, on obtient

$$\Delta_4 = (-1)^6 \begin{vmatrix} -15 & 6 & 10 \\ 22 & -22 & -25 \\ -9 & 2 & 7 \end{vmatrix} = 2 \begin{vmatrix} -15 & 3 & 10 \\ 22 & -11 & -25 \\ -9 & \boxed{1} & 7 \end{vmatrix} =$$

$$= 2\cdot(-1)^{3+2} \begin{vmatrix} 12 & -11 \\ -77 & 52 \end{vmatrix} = 446.$$

Constatons que le nombre de multiplications et de divisions nécessaires pour calculer le déterminant d'ordre n est [8]

$$\frac{n-1}{3}(n^2+n+3).$$

BIBLIOGRAPHIE

1. *O. Schreier, E. Schperner.* Théorie des matrices. ONTI, 1936, §§ 1, 2.
2. *A. Maltsev.* Principes d'algèbre linéaire. Ed. 2, Gostekhizdat, 1956.
3. *V. Faddéeva.* Méthodes numériques d'algèbre linéaire. Gostekhizdat, 1950.
4. *R. A. Frazer, W. J. Duncan, A. R. Collar.* Elementary matrices and some applications to dynamics and differential equations. Cambridge. The Univ. press, 1950.
5. *B. Boulgakov.* Oscillations. Gostekhizdat, 1954, chapitre I.
6. *E. Liapine.* Cours d'algèbre supérieure. Outchpedguiz, 1953, chapitre IX.
7. *E. Whittaker, G. Robinson.* The calculus of observations. A treatise on numerical mathematics. Blackie and Son Ltd., London and Glasgow, 1944.
8. *D. Faddéev, V. Faddéeva.* Méthodes numériques de l'algèbre linéaire. Fizmatguiz, 1960, chapitre II.

CHAPITRE VIII

SYSTÈMES D'ÉQUATIONS LINÉAIRES

§ 1. Généralités sur les méthodes de résolution

Les méthodes de résolution des systèmes d'équations linéaires peuvent être rangées en principe en deux groupes: 1) *méthodes exactes* qui sont des algorithmes finis de calcul des solutions du système (règle de Cramer, méthode de Gauss, méthode du pivot, méthode des racines carrées, etc.) et 2) *méthodes itératives* qui permettent d'obtenir les solutions des systèmes avec la précision imposée à l'aide des processus convergents infinis (*méthode des approximations successives*, *méthode de Seidel*, *méthode de relaxation*, etc.).

Les approximations inévitables font que même les résultats des méthodes exactes sont a p p r o c h é s et, dans le cas général, l'estimation des erreurs des solutions est plutôt malaisée. L'utilisation des processus itératifs donne lieu, en outre, à l'erreur de la méthode.

Remarquons que l'efficacité de l'application des méthodes itératives dépend dans une grande mesure du bon choix de l'approximation initiale et de la rapidité de la convergence du processus.

§ 2. Application de la matrice inverse à la résolution des systèmes. Formules de Cramer

Soit le système de n équations linéaires à n inconnues

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{array}\right\} \quad (1)$$

Désignons par

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \quad (2)$$

la matrice des coefficients du système (1), par

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{3}$$

la colonne de ses termes constants et par

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{4}$$

la colonne des inconnues (v e c t e u r r e c h e r c h é). Alors le système (1) peut être écrit en abrégé sous forme d'une équation matricielle

$$Ax = b. \tag{5}$$

L'ensemble des nombres $x_1, x_2, \ldots, x_n$ (ou tout simplement le vecteur x) qui transforme le système (1) en une identité s'appelle *solution* de ce système et les nombres x_i eux-mêmes, ses *racines.*

Si la matrice A est régulière, c'est-à-dire si

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} = \Delta \neq 0, \tag{6}$$

le système (1) ou l'équation matricielle équivalente (5) possède une solution unique.

En effet, sous la condition $\det A \neq 0$, il existe une matrice inverse A^{-1}. En multipliant les deux membres de l'équation (5) à gauche par la matrice A^{-1}, on obtient

$$A^{-1}Ax = A^{-1}b$$

ou

$$x = A^{-1}b. \tag{7}$$

Il est clair que la formule (7) fournit la solution de l'équation (5) et cette solution est unique, puisque chaque solution est de la forme (7).

E x e m p l e 1. Résoudre le système d'équations

$$\left.\begin{aligned} 3x_1 - x_2 &= 5, \\ -2x_1 + x_2 + x_3 &= 0, \\ 2x_1 - x_2 + 4x_3 &= 15. \end{aligned}\right\}$$

Solution. Mettons le système sous une forme matricielle

$$\begin{bmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \\ 15 \end{bmatrix}.$$

Le déterminant de la matrice A du système considéré

$$\det A = \begin{vmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{vmatrix} = 5 \neq 0.$$

Calculant la matrice inverse A^{-1} on obtient:

$$A^{-1} = \begin{bmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{bmatrix}.$$

D'où

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & \frac{4}{5} & -\frac{1}{5} \\ 2 & \frac{12}{5} & -\frac{3}{5} \\ 0 & \frac{1}{5} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 5 \\ 0 \\ 15 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}.$$

Donc $x_1 = 2$; $x_2 = 1$; $x_3 = 3$.

La recherche directe de l'inverse A^{-1} de la matrice A d'ordre $n > 4$ demande beaucoup de temps. C'est pourquoi il est rare que la formule (7) soit pratiquement employée.

Utilisant la formule (7) il est facile d'obtenir les formules des inconnues du système (1). On sait que (chapitre VII, § 4)

$$A^{-1} = \frac{1}{\Delta} \tilde{A},$$

où

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & \ldots & A_{n1} \\ \ldots & \ldots & \ldots & \ldots \\ A_{1n} & A_{2n} & \ldots & A_{nn} \end{bmatrix}$$

est la matrice adjointe de A (les A_{ij} sont les cofacteurs des éléments a_{ij}). Donc

$$\boldsymbol{x} = \frac{1}{\Delta} \tilde{A} \boldsymbol{b}$$

ou

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_n \end{bmatrix}, \tag{8}$$

avec

$$\Delta_i = \sum_{j=1}^{n} A_{ji} b_j = \begin{vmatrix} a_{11} & \ldots & a_{1,\, i-1} b_1 a_{1,\, i+1} & \ldots & a_{1n} \\ a_{21} & \ldots & a_{2,\, i-1} b_2 a_{2,\, i+1} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \ldots & a_{n,\, i-1} b_n a_{n,\, i+1} & \ldots & a_{nn} \end{vmatrix}$$

qui sont les déterminants déduits du déterminant Δ [formule (6)] en substituant à son i-ième colonne la colonne des termes constants du système (1). L'égalité (8) conduit aux *formules de Cramer*

$$x_1 = \frac{\Delta_1}{\Delta}, \quad x_2 = \frac{\Delta_2}{\Delta}, \quad \ldots, \quad x_n = \frac{\Delta_n}{\Delta}. \tag{9}$$

Donc si le déterminant du système (1) $\Delta \neq 0$, le système possède une solution unique x définie par la formule matricielle (7) ou par des formules scalaires équivalentes (9).

Exemple 2. Résoudre le système d'équations linéaires

$$\left.\begin{aligned} 2x_1 + x_2 - 5x_3 + x_4 &= 8, \\ x_1 - 3x_2 - 6x_4 &= 9, \\ 2x_2 - x_3 + 2x_4 &= -5, \\ x_1 + 4x_2 - 7x_3 + 6x_4 &= 0. \end{aligned}\right\}$$

Solution. Le déterminant de ce système

$$\Delta = \begin{vmatrix} 2 & 1 & -5 & 1 \\ 1 & -3 & 0 & -6 \\ 0 & 2 & -1 & 2 \\ 1 & 4 & -7 & 6 \end{vmatrix} = 27 \neq 0.$$

En calculant les déterminants supplémentaires on obtient :

$$\Delta_1 = \begin{vmatrix} 8 & 1 & -5 & 1 \\ 9 & -3 & 0 & -6 \\ -5 & 2 & -1 & 2 \\ 0 & 4 & -7 & 6 \end{vmatrix} = 81;$$

$$\Delta_2 = \begin{vmatrix} 2 & 8 & -5 & 1 \\ 1 & 9 & 0 & -6 \\ 0 & -5 & -1 & 2 \\ 1 & 0 & -7 & 6 \end{vmatrix} = -108;$$

$$\Delta_3 = \begin{vmatrix} 2 & 1 & 8 & 1 \\ 1 & -3 & 9 & -6 \\ 0 & 2 & -5 & 2 \\ 1 & 4 & 0 & 6 \end{vmatrix} = -27;$$

$$\Delta_4 = \begin{vmatrix} 2 & 1 & -5 & 8 \\ 1 & -3 & 0 & 9 \\ 0 & 2 & -1 & -5 \\ 1 & 4 & -7 & 0 \end{vmatrix} = 27.$$

D'où

$$x_1 = \frac{\Delta_1}{\Delta} = \frac{81}{27} = 3;$$

$$x_2 = \frac{\Delta_2}{\Delta} = -\frac{108}{27} = -4;$$

$$x_3 = \frac{\Delta_3}{\Delta} = -\frac{27}{27} = -1;$$

$$x_4 = \frac{\Delta_4}{\Delta} = \frac{27}{27} = 1.$$

La résolution du système linéaire (1) à n inconnues se ramène ainsi au calcul de $(n+1)$-ième déterminant d'ordre n. Si le nombre n est grand, le calcul des déterminants est une opération délicate. Aussi pour le calcul des racines d'un système linéaire a-t-on établi des procédés directs.

§ 3. Méthode de Gauss

La méthode la plus usitée de résolution des systèmes d'équations linéaires est l'algorithme d'élimination successive des inconnues. Cette méthode s'appelle *méthode de Gauss*. Pour simplifier les raisonnements, bornons-nous à considérer un système de quatre équations à quatre inconnues

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = a_{15}, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = a_{25}, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = a_{35}, \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 = a_{45}. \end{aligned}\right\} \quad (1)$$

Soit $a_{11} \neq 0$ (élément générateur). Divisant les coefficients de la première équation du système (1) par a_{11}, on obtient

$$x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 = b_{15}, \quad (2)$$

où

$$b_{1j} = \frac{a_{1j}}{a_{11}} \qquad (j > 1).$$

En appliquant l'équation (2) on élimine facilement l'inconnue x_1 du système (1). A cette fin il suffit de soustraire de la deuxième équation de (1) le produit de l'équation (2) par a_{21}, de la troisième équation de (1) le produit de l'équation (2) par a_{31}, etc. Il en résulte un système de trois équations

$$\left.\begin{aligned} a_{22}^{(1)}x_2+a_{23}^{(1)}x_3+a_{24}^{(1)}x_4&=a_{25}^{(1)},\\ a_{32}^{(1)}x_2+a_{33}^{(1)}x_3+a_{34}^{(1)}x_4&=a_{35}^{(1)},\\ a_{42}^{(1)}x_2+a_{43}^{(1)}x_3+a_{44}^{(1)}x_4&=a_{45}^{(1)},\end{aligned}\right\} \qquad (1')$$

dont les coefficients $a_{ij}^{(1)}$ $(i, j \geqslant 2)$ se calculent d'après la formule

$$a_{ij}^{(1)}=a_{ij}-a_{i1}b_{1j} \quad (i, j \geqslant 2).$$

Après avoir divisé ensuite les coefficients de la première équation du système (1′) par l'« élément générateur » $a_{22}^{(1)}$, on a l'équation

$$x_2+b_{23}^{(1)}x_3+b_{24}^{(1)}x_4=b_{25}^{(1)}, \qquad (2')$$

où

$$b_{2j}^{(1)}=\frac{a_{2j}^{(1)}}{a_{22}^{(1)}} \quad (j>2).$$

Eliminons maintenant x_2 de la même façon que x_1 pour aboutir au système:

$$\left.\begin{aligned} a_{33}^{(2)}x_3+a_{34}^{(2)}x_4&=a_{35}^{(2)},\\ a_{43}^{(2)}x_3+a_{44}^{(2)}x_4&=a_{45}^{(2)},\end{aligned}\right\} \qquad (1'')$$

où

$$a_{ij}^{(2)}=a_{ij}^{(1)}-a_{i2}^{(1)}b_{2j}^{(1)} \quad (i, j \geqslant 3).$$

Les coefficients de la première équation de (1″) divisés par l'« élément générateur » $a_{33}^{(2)}$ donnent

$$x_3+b_{34}^{(2)}x_4=b_{35}^{(2)}, \qquad (2'')$$

où

$$b_{3j}^{(2)}=\frac{a_{3j}^{(2)}}{a_{33}^{(2)}} \quad (j>3).$$

En éliminant maintenant d'une façon analogue x_3 du système (1″), on obtient:

$$a_{44}^{(3)}x_4=a_{45}^{(3)}, \qquad (1''')$$

où

$$a_{ij}^{(3)}=a_{ij}^{(2)}-a_{i3}^{(2)}b_{3j}^{(2)} \quad (i, j \geqslant 4).$$

D'où

$$x_4 = \frac{a_{45}^{(3)}}{a_{44}^{(3)}} = b_{45}^{(3)}. \qquad (2''')$$

Les autres inconnues sont données successivement par les équations (2''), (2') et (2):

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4,$$
$$x_2 = b_{25}^{(1)} - b_{24}^{(1)} x_4 - b_{23}^{(1)} x_3,$$
$$x_1 = b_{15} - b_{14} x_4 - b_{13} x_3 - b_{12} x_2.$$

Ainsi la procédure de résolution d'un système linéaire d'après la méthode de Gauss se ramène à la construction d'un système équivalent (2), (2'), (2''), (2''') à matrice triangulaire. La condition nécessaire et suffisante pour l'application de la méthode est que tous les « éléments générateurs » soient non nuls. Il est commode de ranger les résultats des calculs dans le tableau 13. Le schéma donné par ce tableau s'appelle *schéma de division unique*. La procédure de recherche des coefficients $b_{ij}^{(j-1)}$ du système triangulaire s'appelle dans le cas général *marche directe*, celle d'obtention des valeurs des inconnues, *marche inverse*.

La marche directe débute par l'inscription des coefficients du système, y compris des termes constants (section A). Sur la dernière ligne de la section A figure le résultat de la division de la première ligne de la section par l'« élément générateur » a_{11}. Les éléments $a_{ij}^{(1)}$ (i, $j \geqslant 2$) de la section suivante A_1 du schéma sont égaux aux éléments correspondants a_{ij} de la section précédente diminués du produit de leurs « projections » par les colonnes de la section A qui portent l'élément 1 (c'est-à-dire par la première colonne et la dernière ligne).

La dernière ligne de la section A_1 s'obtient en divisant la première ligne de la section par l'« élément générateur » $a_{22}^{(1)}$. D'une façon analogue on construit les sections suivantes. La marche directe s'arrête lorsqu'on atteint la section composée d'une ligne, sans compter la ligne transformée (dans le cas concerné c'est la section A_3).

La marche inverse ne fait appel qu'aux lignes des sections A_i qui contiennent les unités (*lignes marquées*) en commençant par la dernière. L'élément $b_{45}^{(3)}$ de la section A_3 figurant à l'intersection de la colonne des termes constants et de la ligne marquée de la section donne la valeur de x_4. Ensuite, toutes les autres inconnues x_i ($i = 3, 2, 1$) se trouvent de proche en proche en retranchant du terme constant de la ligne marquée la somme des produits de ses coefficients par les valeurs correspondantes des inconnues trouvées auparavant. Les valeurs des inconnues sont portées successivement sur la dernière section B. Les unités qui y figurent aident à trouver pour x_i les coefficients respectifs dans les lignes marquées.

Tableau 13

Schéma de division unique

x_1	x_2	x_3	x_4	Termes constants	Σ	Sections du schéma
a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	
a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	
a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	A
a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	
1	b_{12}	b_{13}	b_{14}	b_{15}	b_{16}	
	$a_{22}^{(1)}$	$a_{23}^{(1)}$	$a_{24}^{(1)}$	$a_{25}^{(1)}$	$a_{26}^{(1)}$	
	$a_{32}^{(1)}$	$a_{33}^{(1)}$	$a_{34}^{(1)}$	$a_{35}^{(1)}$	$a_{36}^{(1)}$	A_1
	$a_{42}^{(1)}$	$a_{43}^{(1)}$	$a_{44}^{(1)}$	$a_{45}^{(1)}$	$a_{46}^{(1)}$	
	1	$b_{23}^{(1)}$	$b_{24}^{(1)}$	$b_{25}^{(1)}$	$b_{26}^{(1)}$	
		$a_{33}^{(2)}$	$a_{34}^{(2)}$	$a_{35}^{(2)}$	$a_{36}^{(2)}$	
		$a_{43}^{(2)}$	$a_{44}^{(2)}$	$a_{45}^{(2)}$	$a_{46}^{(2)}$	A_2
		1	$b_{34}^{(2)}$	$b_{35}^{(2)}$	$b_{36}^{(2)}$	
			$a_{44}^{(3)}$	$a_{45}^{(3)}$	$a_{46}^{(3)}$	
			1	$b_{45}^{(3)}$	$b_{46}^{(3)}$	A_3
				(x_4)	$(\bar{x}_4)$	
			1	x_4	$\bar{x}_4$	
		1		x_3	$\bar{x}_3$	B
	1			x_2	$\bar{x}_2$	
1				x_1	$\bar{x}_1$	

Pour vérifier les calculs on utilise ce qu'on appelle les « sommes de contrôle »

$$a_{i6}=\sum_{j=1}^{5} a_{ij} \quad (i=1, 2, \ldots, 5), \tag{3}$$

portées sur la colonne $\sum$ et qui constituent la somme des éléments des lignes de la matrice du système initial (1), y compris les termes constants.

Admettons que a_{i6} sont les nouveaux termes constants du système (1), alors le système linéaire transformé

$$\sum_{j=1}^{4} a_{ij}\bar{x}_j = a_{i6} \quad (i=1, 2, 3, 4) \tag{4}$$

aura des inconnues $\bar{x}_j$ associées aux inconnues précédentes x_j par les relations

$$\bar{x}_j = x_j + 1 \quad (j = 1, 2, 3, 4). \tag{5}$$

En effet, en portant les formules (5) dans l'équation (4) on obtient, en vertu du système (1) et des formules (3), l'identité

$$\sum_{j=1}^{4} a_{ij}x_j + \sum_{j=1}^{4} a_{ij} = \sum_{j=1}^{5} a_{ij} \equiv a_{i6} \quad (j = 1, 2, 3, 4).$$

En général, si on effectue sur les sommes de contrôle dans chaque ligne les mêmes opérations que sur tous les autres éléments de cette ligne, en l'absence d'erreurs de calcul, les éléments de la colonne Σ sont égaux aux sommes des éléments des lignes transformées correspondantes, ce qui permet de vérifier la marche directe. La marche inverse est vérifiée par la recherche des nombres $\bar{x}_j$ qui doivent coïncider avec les nombres $x_j + 1$.

Exemple. Résoudre le système

$$\left.\begin{aligned} 7,9x_1 + 5,6x_2 + 5,7x_3 - 7,2x_4 &= 6,68\,; \\ 8,5x_1 - 4,8x_2 + 0,8x_3 + 3,5x_4 &= 9,95\,; \\ 4,3x_1 + 4,2x_2 - 3,2x_3 + 9,3x_4 &= 8,6\,; \\ 3,2x_1 - 1,4x_2 - 8,9x_3 + 3,3x_4 &= 1. \end{aligned}\right\} \tag{6}$$

Solution. Portons sur la section A du tableau 14 la matrice des coefficients du système, ses termes constants et les sommes de contrôle. Inscrivons ensuite la dernière (cinquième) ligne de la section A en divisant la première ligne par 7,9 (par a_{11}).

Passons maintenant à la section A_1 du tableau. Prenons un élément quelconque de la section A (absent dans la première ligne) et retranchons le produit du premier élément de sa ligne par le dernier élément de sa colonne pour inscrire le résultat dans la case correspondante de la section A_1 du schéma. Par exemple, en choisissant $a_{43} = -8,9$, on aura:

$$a_{43}^{(1)} = a_{43} - a_{41}b_{13} = -8,9 - 3,2 \cdot 0,72152 = -11,20886.$$

Pour obtenir la dernière ligne de la section A_1, divisons tous les éléments de la première ligne de cette section par $a_{22}^{(1)} = -10,82531$. Par exemple,

$$b_{231}^{(1)} = \frac{a_{23}^{(1)}}{a_{22}^{(1)}} = \frac{-5,33292}{-10,82531} = 0,49263.$$

On remplit d'une façon analogue les autres sections du tableau. Par exemple,

$$a_{44}^{(2)} = a_{44}^{(1)} - a_{42}^{(1)}b_{24}^{(1)} = 6,21645 - (-3,66835)\cdot(-1,03894) = 2,40525.$$

Tableau 14

Résolution d'un système d'après le schéma de division unique

x_1	x_2	x_3	x_4	Termes constants	Σ	Sections du schéma
7,9	5,6	5,7	−7,2	6,68	18,68	*A*
8,5	−4,8	0,8	3,5	9,95	17,95	
4,3	4,2	−3,2	9,3	8,6	23,2	
3,2	−1,4	−8,9	3,3	1	−2,8	
1	0,70886	0,72152	−0,91139	0,84557	2,36456	
	−10,82531	−5,33292	11,24682	2,76265	−2,14876	A_1
	1,15190	−6,30254	13,21898	4,96405	13,03239	
	−3,66835	−11,20886	6,21645	−1,70582	−10,36658	
	1	0,49263	−1,03894	−0,25520	0,19849	
		−6,87000	14,41573	5,25801	12,80374	A_2
		−9,40172	2,40525	−2,64198	−9,63845	
		1	−2,09836	−0,76536	−1,86372	
			−17,32294	−9,83768	−27,16062	A_3
			1	0,56790	1,56790	
			1	0,56790	1,56790	*B*
		1		0,42630	1,42630	
	1			0,12480	1,12480	
1				0,96710	1,96710	

Pour trouver les inconnues considérons les lignes contenant les unités (*lignes marquées*) en commençant par la dernière. L'inconnue x_4 est le terme constant de la dernière ligne de la section A_3:

$$x_4 = b_{45}^{(3)} = 0{,}56790.$$

Les valeurs des autres inconnues x_3, x_2, x_1 s'obtiennent successivement en retranchant des termes constants figurant sur les lignes marquées la somme des produits des coefficients respectifs $b_{ij}^{(k)}$ par les valeurs des inconnues trouvées auparavant.

On a:

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4 = -0{,}76536 - (-2{,}09836) \cdot 0{,}56790 = 0{,}42630;$$

$$x_2 = b_{25}^{(1)} - b_{24}^{(1)} x_4 - b_{23}^{(1)} x_3 =$$
$$= -0{,}25520 - (-1{,}03894) \cdot 0{,}56790 - 0{,}49263 \cdot 0{,}42630 = 0{,}12480;$$

$$x_1 = b_{15} - b_{14} x_4 - b_{13} x_3 - b_{12} x_2 = 0{,}84557 - (-0{,}91139) \cdot 0{,}56790 -$$
$$- 0{,}72152 \cdot 0{,}42630 - 0{,}70886 \cdot 0{,}12480 = 0{,}96710.$$

Ainsi

$$x_1 = 0{,}96710; \quad x_2 = 0{,}12480; \quad x_3 = 0{,}42630; \quad x_4 = 0{,}56790.$$

La vérification courante des calculs s'effectue à l'aide de la colonne Σ soumise aux mêmes opérations que les autres colonnes.

Il en résulte que 1) la somme des éléments de chaque ligne du schéma (absents dans la colonne Σ) doit être égale à l'élément de cette ligne figurant dans la colonne Σ ; 2) les nombres $\overline{x}_i$ dans la colonne Σ doivent être supérieurs d'une unité aux racines respectives de la solution du système.

A propos, si l'on tient compte des unités figurant dans la section B, on obtient encore que dans cette section les éléments de la colonne Σ sont les sommes des éléments des lignes qui leur correspondent. Dans le cas concerné, la première et la deuxième condition sont observées à une unité du dernier rang près. Par conséquent, il est presque certain que les calculs sont corrects.

Constatons que si la matrice du système est symétrique, les parties respectives des sections A, A_1, A_2, . . . du schéma de division unique sont symétriques elles aussi. Cette circonstance peut être mise à profit pour simplifier le tableau.

L'estimation du nombre N d'opérations arithmétiques nécessaires pour résoudre un système linéaire à n inconnues par la méthode de Gauss [5] (sans tenir compte de la vérification) ne présente aucune difficulté. Le nombre de multiplications et de divisions nécessaires pour la marche directe est

$$n(n+1)+(n-1)n+\ldots+1\cdot 2 =$$
$$=(1^2+2^2+\ldots+n^2)+(1+2+\ldots+n)=\frac{n(n+1)(n+2)}{3},$$

c'est aussi le nombre de soustractions. Pour la marche inverse il faut $\frac{n(n-1)}{2}$ multiplications et divisions et le même nombre de soustractions. Avec $n > 7$, le nombre total d'opérations arithmétiques imposées par la méthode de Gauss est donc

$$N=\frac{2n(n+1)(n+2)}{3}+n(n-1)<n^3.$$

Ainsi le temps nécessaire pour résoudre un système linéaire par la méthode de Gauss est à peu près proportionnel au cube du nombre d'inconnues. Par exemple, pour résoudre un système de 100 équations linéaires à 100 inconnues sur une machine rapide qui effectue 10^4 opérations par seconde, il faut

$$T = 10^6 \cdot 10^{-4} = 100 \text{ s}.$$

Le temps machine réel sera beaucoup plus grand par suite de la présence dans le programme d'opérations autres que les opérations arithmétiques (substitution d'adresse, opérations logiques, transferts, mise en forme, etc.).

§ 4. Amélioration de la précision des racines

Les solutions approchées obtenues par la méthode de Gauss peuvent être précisées. Montrons comment il faut procéder à cet effet si les corrections des racines sont petites en valeur absolue.

Supposons qu'on ait trouvé pour le système

$$Ax = b$$

la solution approchée x_0. Posant

$$x = x_0 + \delta,$$

pour la correction $\delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_n \end{bmatrix}$ de la solution x_0, on a l'équation

$$A(x_0 + \delta) = b$$

ou

$$A\delta = \varepsilon,$$

$\varepsilon = b - Ax_0$ étant le *résidu de la solution approchée* x_0. Ainsi pour obtenir δ, il faut résoudre le système linéaire à matrice précédente A et au nouveau terme constant $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$. A cet effet il suffit d'ajouter au schéma de calcul principal la colonne ε des termes constants et la transformer d'après les règles générales. Suivant l'usage, les corrections $\delta_1, \delta_2, \ldots, \delta_n$ sont déterminées à partir des lignes marquées, les coefficients de ces corrections inconnues étant déjà fournis par le tableau. Notons qu'on peut ne pas préciser les coefficients transformés de la matrice A, car dans le cas de faibles résidus l'ordre de petitesse des erreurs respectives est plus grand.

E x e m p l e. Résoudre par la méthode de Gauss avec trois chiffres (avec une règle de calcul, par exemple, ou à la main) le système

$$\left.\begin{aligned} 6x_1 - x_2 - x_3 &= 11{,}33\,; \\ -x_1 + 6x_2 - x_3 &= 32\,; \\ -x_1 - x_2 + 6x_3 &= 42. \end{aligned}\right\} \qquad (1)$$

En utilisant les valeurs obtenues comme des approximations initiales, améliorer la précision des racines jusqu'à 10^{-4}.

S o l u t i o n. En appliquant le schéma usuel de la division unique (tableau 15), on effectue toutes les opérations avec trois chiffres significatifs.

On a les solutions approchées:

$$x_1^{(0)} = 4{,}67\,; \quad x_2^{(0)} = 7{,}62\,; \quad x_3^{(0)} = 9{,}05.$$

En portant ces nombres dans le système (1), on calcule les résidus correspondants (c'est-à-dire les différences entre le premier et le deuxième membre du système (1))

$$\varepsilon_1^{(0)} = -0{,}02\,; \quad \varepsilon_2^{(0)} = 0\,; \quad \varepsilon_3^{(0)} = -0{,}01.$$

Tableau 15

Amélioration de la précision des solutions calculées par la méthode de Gauss

x_1	x_2	x_3	Termes constants	Σ	Résidu ε
6 —1 —1	—1 6 —1	—1 —1 6	11,33 32 42	15,33 36 46	—0,02 0 —0,01
1	—0,167	—0,167	1,89	2,56	—0,0033
	5,83 —1,17	—1,17 5,83	33,9 43,9	38,6 48,6	—0,0033 —0,0133
	1	—0,200	5,80	6,60	—0,0006
		5,60	50,7	56,3	—0,0140
		1	9,05 9,0475	10,05	—0,0025
	1		7,62 7,6189	8,62	—0,0011
1			4,67 4,6661		—0,0039

Utilisons ces valeurs comme termes constants (tableau 15) pour obtenir la correction des solutions

$$\delta_1^{(0)} = -0{,}0039\,; \quad \delta_2^{(0)} = -0{,}0011\,; \quad \delta_3^{(0)} = -0{,}0025.$$

On en tire les solutions précisées

$$x_1 = 4{,}6661\,; \quad x_2 = 7{,}6189\,; \quad x_3 = 9{,}0475,$$

les résidus étant égaux à

$$\delta_1 = -2\cdot 10^{-4}; \quad \delta_2 = 2\cdot 10^{-4}; \quad \delta_3 = 0.$$

Parfois il faut déterminer une erreur éventuelle Δx de la solution x d'un système linéaire d'après les petites erreurs connues ΔA et Δb de la matrice A du système et de son terme constant b.

On a:

$$Ax = b \tag{2}$$

et

$$(A + \Delta A)(x + \Delta x) = b + \Delta b.$$

Il en résulte si l'on néglige le petit terme $\Delta A \cdot \Delta x$

$$Ax + A\Delta x + \Delta A x = b + \Delta b \tag{3}$$

ou

$$A\Delta x = \Delta b - \Delta A x. \qquad (4)$$

Ainsi, lors de la recherche approchée de Δx, on peut utiliser le schéma de Gauss pour le système principal (2) en complétant ce schéma par une nouvelle colonne de termes constants $\Delta b - \Delta A x$.

§ 5. Méthode du pivot

Soit le système linéaire

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = a_{1.\,n+1}, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = a_{2.\,n+1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = a_{n.\,n+1}. \end{array}\right\} \qquad (1)$$

Considérons la matrice rectangulaire étendue des coefficients du système et de ses termes constants

$$M = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1j} & \ldots & a_{1q} & \ldots & a_{1n} & a_{1.\,n+1} \\ a_{21} & a_{22} & \ldots & a_{2j} & \ldots & a_{2q} & \ldots & a_{2n} & a_{2.\,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{i1} & a_{i2} & \ldots & a_{ij} & \ldots & a_{iq} & \ldots & a_{in} & a_{i.\,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{p1} & a_{p2} & \ldots & a_{pj} & \ldots & \boxed{a_{pq}} & \ldots & a_{pn} & a_{p.\,n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nj} & \ldots & a_{nq} & \ldots & a_{nn} & a_{n.\,n+1} \end{bmatrix}.$$

Choisissons l'élément non nul a_{pq} de la matrice M en général le plus grand en module et n'appartenant pas à la colonne des termes constants ($q \neq n + 1$) qu'on appelle *pivot* et calculons les facteurs

$$m_i = -\frac{a_{iq}}{a_{pq}}$$

pour tout $i \neq p$.

La ligne affectée de numéro p de la matrice M, qui contient le pivot, s'appelle *ligne du pivot*. En poursuivant on effectue l'opération suivante: ajoutons à chaque ligne sans pivot le produit de la ligne du pivot par le facteur correspondant m_i de cette ligne. Il en résulte une nouvelle matrice dont la q-ième colonne est composée de zéros. En rejetant cette colonne et la p-ième ligne du pivot, on obtient une nouvelle matrice $M^{(1)}$ avec le nombre de lignes et de colonnes diminué d'une unité.

Reprenant les mêmes opérations avec la matrice $M^{(1)}$, on obtient la matrice $M^{(2)}$, etc. On construit ainsi une suite des matrices

$$M,\ M^{(1)},\ \ldots,\ M^{(n-1)},$$

dont la dernière est une matrice ligne à deux termes; considérons-la également comme ligne du pivot.

Pour déterminer les inconnues x_i associons en un système toutes les lignes du pivot à partir de la dernière qui appartient à la matrice $M^{(n-1)}$.

Après avoir dûment changé la numérotation des inconnues, on obtient un système à matrice triangulaire qui permet de calculer sans peine de proche en proche les inconnues du système (1) donné. La méthode du pivot peut toujours être appliquée si le déterminant du système

$$\det A = \begin{vmatrix} a_{11} & \ldots & a_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{n1} & \ldots & a_{nn} \end{vmatrix} \neq 0.$$

Notons que la méthode de Gauss est un cas particulier de la méthode du pivot et le schéma de la méthode de Gauss s'obtient si l'on choisit toujours comme pivot l'élément supérieur gauche de la matrice correspondante.

§ 6. Application de la méthode de Gauss au calcul des déterminants

Soit

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \tag{1}$$

et

$$\Delta = \det A. \tag{2}$$

Considérons le système linéaire

$$Ax = 0. \tag{3}$$

En résolvant le système (3) par la méthode de Gauss, nous avons remplacé la matrice A par la matrice triangulaire B composée d'éléments des lignes marquées

$$B = \begin{bmatrix} 1 & b_{12} & b_{13} & \ldots & b_{1n} \\ 0 & 1 & b_{23}^{(1)} & \ldots & b_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

Il en a résulté un système équivalent

$$Bx = 0. \tag{4}$$

Les éléments de la matrice B s'obtenaient successivement à partir des éléments de la matrice A et des matrices auxiliaires ultérieures $A_1, A_2, \ldots, A_{n-1}$ à l'aide des transformations élémentaires suivantes:

1) division par les éléments « générateurs » $a_{11}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$ supposés différents de zéro, et

2) soustraction aux lignes de la matrice A et aux matrices intermédiaires A_i ($i = 1, 2, \ldots, n-1$) des nombres proportionnels aux éléments des lignes génératrices correspondantes. Dans la première opération le déterminant de la matrice est de même divisé par l'élément « générateur » correspondant, dans la deuxième ce déterminant reste inchangé. C'est pourquoi

$$\det B = 1 = \frac{\det A}{a_{11}a_{22}^{(1)} \ldots a_{nn}^{(n-1)}}.$$

Par conséquent,

$$\Delta = \det A = a_{11}a_{22}^{(1)} \ldots a_{nn}^{(n-1)}, \tag{5}$$

c'est-à-dire le *déterminant est égal au produit des éléments « générateurs » du schéma de Gauss correspondant.* On en déduit que le schéma de division unique du § 3 peut être utilisé pour le calcul des déterminants en rejetant la colonne des termes constants comme inutile.

Notons que si à une étape quelconque l'élément $a_{ii}^{(i-1)} = 0$ ou voisin de zéro (ce qui entraîne l'altération de la précision du calcul), il convient de réaliser la permutation appropriée des lignes et des colonnes de la matrice.

E x e m p l e. Calculer le déterminant

$$\Delta = \begin{vmatrix} 7,4 & 2,2 & -3,1 & 0,7 \\ 1,6 & 4,8 & -8,5 & 4,5 \\ 4,7 & 7,0 & -6,0 & 6,6 \\ 5,9 & 2,7 & 4,9 & -5,3 \end{vmatrix}.$$

S o l u t i o n. Utilisons les éléments du déterminant Δ pour composer le schéma de division unique (tableau 16).

En multipliant entre eux les éléments « générateurs » (encadrés) on obtient:

$$\Delta = 7,4 \cdot 4,32434 \cdot 6,11331 \cdot (-7,58393) = -1483,61867.$$

Insistons sur la circonstance suivante. Pour résoudre un système de n équations linéaires à n inconnues d'après les formules de Cramer, il faut calculer $n + 1$ déterminants d'ordre n. Or pour cal-

Tableau 16

Calcul du déterminant par la méthode de Gauss

1-ère colonne	2-e colonne	3-e colonne	4-e colonne	Σ	
7,4	2,2	−3,1	0,7	7,2	
1,6	4,8	−8,5	4,5	2,4	
4,7	7,0	−6,0	6,6	12,3	A
5,9	2,7	4,9	−5,3	8,2	
1	0,29729	−0,41891	0,09459	0,97297	
	4,32434	−7,82974	4,34866	0,84326	
	5,60274	−4,03112	6,15543	7,72705	A_1
	0,94599	7,37157	−5,85808	2,45948	
	1	−1,81062	1,00562	0,19500	
		6,11331	0,52120	6,63451	
		9,08440	−6,80939	2,27501	A_2
		1	0,08526	1,08526	
			−7,58393	−7,58393	A_3
				$\Delta = -1483,61867$	

culer un seul déterminant d'ordre n suivant le schéma de division unique il faut réaliser presque le même travail que dans le cas de la résolution complète d'un système d'équations. Donc, en général, dans le cas de calcul numérique d'un système linéaire avec $n > 3$, les formules de Cramer ne présentent aucun avantage.

§ 7. Calcul d'une matrice inverse par la méthode de Gauss

Soit la matrice régulière

$$A = [a_{ij}] \quad (i, j = 1, 2, \ldots, n). \tag{1}$$

Pour trouver son inverse

$$A^{-1} = [x_{ij}] \tag{2}$$

on utilise la relation principale

$$AA^{-1} = E, \tag{3}$$

où E est une matrice unité.

Tableau 17

Calcul de la matrice inverse par la méthode de Gauss

x_{1j}	x_{2j}	x_{3j}	x_{4j}	$j=1$	$j=2$	$j=3$	$j=4$	Σ
1,8	—3,8	0,7	—3,7	1	0	0	0	—4,0
0,7	2,1	—2,6	—2,8	0	1	0	0	—1,6
7,3	8,1	1,7	—4,9	0	0	1	0	13,2
1,9	—4,3	—4,9	—4,7	0	0	0	1	—11,0
1	—2,11111	0,38889	—2,05556	0,55556	0	0	0	—2,22223
	3,57778	—2,87222	—1,36111	—0,38885	1	0	0	—0,04440
	23,51110	—1,13890	10,10559	—4,05551	0	1	0	29,42228
	—0,28889	—5,63889	—0,79444	—1,05554	0	0	1	—6,77776
	1	—0,80279	—0,38043	—0,10868	0,27950	0	0	—0,01241
		17,73557	19,04992	—1,50032	—6,57135	1	0	29,71405
		—5,87081	—0,90434	—1,08694	0,08074	0	1	—6,78134
		1	1,07411	—0,08459	—0,37108	0,05638	0	1,67539
			5,40155	—1,58355	—2,09780	0,33100	1	3,05456
			1	—0,29316	—0,38837	0,06128	0,18513	0,56540
		1		0,23030	0,04607	—0,00944	—0,19885	1,06809
	1			—0,03533	0,16873	0,01573	—0,08920	1,06013
1				—0,21121	—0,46003	0,16284	0,26956	0,76266

En multipliant les matrices A et A^{-1} on obtient n systèmes d'équations par rapport à n^2 inconnues x_{ij}

$$\sum_{k=1}^{n} a_{ik}x_{kj} = \delta_{ij} \quad (i, j = 1, 2, \ldots, n),$$

où

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

Les n systèmes d'équations linéaires obtenus pour $j = 1, 2, \ldots, n$ ayant la même matrice A et des termes constants différents peuvent être résolus simultanément par la méthode de Gauss.

Exemple. Trouver l'inverse A^{-1} de la matrice

$$A = \begin{bmatrix} 1{,}8 & -3{,}8 & 0{,}7 & -3{,}7 \\ 0{,}7 & 2{,}1 & -2{,}6 & -2{,}8 \\ 7{,}3 & 8{,}1 & 1{,}7 & -4{,}9 \\ 1{,}9 & -4{,}3 & -4{,}9 & -4{,}7 \end{bmatrix}.$$

Solution. Composons le schéma de division unique. Nous aurons quatre colonnes de termes constants (tableau 17). Notons que les éléments des lignes de la matrice inverse s'obtiennent dans l'ordre inverse.

Les résultats du tableau 17 conduisent à

$$A^{-1} = \begin{bmatrix} -0{,}21121 & -0{,}46003 & 0{,}16284 & 0{,}26956 \\ -0{,}03533 & 0{,}16873 & 0{,}01573 & -0{,}08920 \\ 0{,}23030 & 0{,}04607 & -0{,}00944 & -0{,}19885 \\ -0{,}29316 & -0{,}38837 & 0{,}06128 & 0{,}18513 \end{bmatrix}.$$

Pour vérifier, composons le produit

$$AA^{-1} = \begin{bmatrix} 1{,}8 & -3{,}8 & 0{,}7 & -3{,}7 \\ 0{,}7 & 2{,}1 & -2{,}6 & -2{,}8 \\ 7{,}3 & 8{,}1 & 1{,}7 & -4{,}9 \\ 1{,}9 & -4{,}3 & -4{,}9 & -4{,}7 \end{bmatrix} \times$$

$$\times \begin{bmatrix} -0{,}21121 & -0{,}46003 & 0{,}16284 & 0{,}26956 \\ -0{,}03533 & 0{,}16873 & 0{,}01573 & -0{,}08920 \\ 0{,}23030 & 0{,}04607 & -0.00944 & -0{,}19885 \\ -0{,}29316 & -0{,}38837 & 0{,}06128 & 0{,}18513 \end{bmatrix} =$$

$$= \begin{bmatrix} 0,99997 & 0,00000 & -0,00001 & 0,00000 \\ -0,00025 & 0,99997 & -0,00002 & -0,00039 \\ -0,00808 & -0,01017 & 0,99982 & 0,00009 \\ 0,00000 & 0,00000 & 0,00000 & 1,00048 \end{bmatrix} =$$

$$= E - 10^{-3} \cdot \begin{bmatrix} 0,03 & 0,00 & 0,01 & 0,00 \\ 0,25 & 0,03 & 0,02 & 0,39 \\ 8,08 & 10,17 & 0,18 & -0,09 \\ 0,00 & 0,00 & 0,00 & -0,48 \end{bmatrix}.$$

On voit que par suite de l'arrondissement la matrice inverse obtenue n'est pas tout à fait exacte. Nous allons indiquer ci-dessous (cf. § 15) une méthode de correction des éléments d'une matrice inverse approchée.

§ 8. Méthode des racines carrées

Soit le système linéaire

$$Ax = b, \tag{1}$$

où $A = [a_{ij}]$ est une matrice symétrique, c'est-à-dire $A' = [a_{ji}] = A$. On peut alors mettre la matrice A sous forme d'un produit de deux matrices triangulaires telles que l'une soit transposée de l'autre:

$$A = T'T, \tag{2}$$

où

$$T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & t_{nn} \end{bmatrix} \quad \text{et} \quad T' = \begin{bmatrix} t_{11} & 0 & \dots & 0 \\ t_{12} & t_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ t_{1n} & t_{2n} & \dots & t_{nn} \end{bmatrix}.$$

Pour trouver les éléments t_{ij} de la matrice T, on obtient les équations suivantes en multipliant les matrices T' et T:

$$\left. \begin{aligned} t_{1i}t_{1j} + t_{2i}t_{2j} + \dots + t_{ii}t_{ij} &= a_{ij} \quad (i < j), \\ t_{1i}^2 + t_{2i}^2 + \dots + t_{ii}^2 &= a_{ii}. \end{aligned} \right\}$$

On en tire successivement:

$$\left. \begin{aligned} & t_{11} = \sqrt{a_{11}}, \quad t_{1j} = \frac{a_{1j}}{t_{11}} && (j > 1), \\ & t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t_{ki}^2} && (1 < i \leqslant n), \\ & t_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} t_{ki}t_{kj}}{t_{ii}} && (i < j), \\ & t_{ij} = 0 \text{ pour } i > j. \end{aligned} \right\} \tag{3}$$

Si $t_{ii} \neq 0$ $(i = 1, 2, \ldots, n)$, le système (1) possède une solution unique définie, du fait que

$$\det A = \det T' \cdot \det T = (\det T)^2 = (t_{11}t_{22} \ldots t_{nn})^2 \neq 0.$$

Les coefficients de la matrice T seront réels si $t_{ii}^2 > 0$. Dans ce qui suit nous ne supposerons pas en général que cette dernière condition est respectée.

Lorsque la relation (2) est vérifiée, l'équation (1) est équivalente aux deux équations

$$T'y = b \quad \text{et} \quad Tx = y,$$

ou sous une forme développée

$$\left.\begin{aligned} &t_{11}y_1 = b_1, \\ &t_{12}y_1 + t_{22}y_2 = b_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &t_{1n}y_1 + t_{2n}y_2 + \ldots + t_{nn}y_n = b_n \end{aligned}\right\} \tag{4}$$

et

$$\left.\begin{aligned} t_{11}x_1 + t_{12}x_2 + \ldots + t_{1n}x_n &= y_1, \\ t_{22}x_2 + \ldots + t_{2n}x_n &= y_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots& \\ t_{nn}x_n &= y_n. \end{aligned}\right\} \tag{5}$$

On en déduit successivement:

$$\left.\begin{aligned} &y_1 = \frac{b_1}{t_{11}}, \\ &y_i = \frac{b_i - \sum\limits_{k=1}^{i-1} t_{ki}y_k}{t_{ii}} \qquad (i > 1) \end{aligned}\right\} \tag{6}$$

et

$$\left.\begin{aligned} &x_n = \frac{y_n}{t_{nn}}, \\ &x_i = \frac{y_i - \sum\limits_{k=i+1}^{n} t_{ik}x_k}{t_{ii}} \qquad (i < n). \end{aligned}\right\} \tag{7}$$

La méthode exposée de résolution d'un système linéaire s'appelle *méthode des racines carrées*. A étant une matrice symétrique et T une matrice triangulaire supérieure, on ne peut inscrire sur le schéma de calcul numérique que $\frac{n}{2}(n + 1)$ coefficients supérieurs a_{ij} et t_{ij} $(i \geqslant j)$. La vérification usuelle se fait à l'aide des sommes, la composition des sommes rendant compte de tous les coefficients de la ligne correspondante.

Remarquons que si pour une certaine s-ième ligne on a $t_{ss}^2 < 0$, les éléments respectifs t_{sj} seront imaginaires. Dans ce cas-là encore la méthode est formellement applicable.

L application pratique de la méthode des racines carrées consiste à calculer successivement par *marche directe* d'après les formules (3) et (6) les coefficients t_{ij} et y_i $(i = 1, 2, \ldots, n)$, puis à calculer par *marche inverse* d'après la formule (7) les inconnues x_i $(i = n, n-1, \ldots, 1)$.

Exemple. Résoudre par la méthode des racines carrées le système d'équations

$$\left.\begin{aligned} x_1 + 3x_2 - 2x_3 \qquad\quad - 2x_5 &= 0,5;\\ 3x_1 + 4x_2 - 5x_3 + x_4 - 3x_5 &= 5,4;\\ -2x_1 - 5x_2 + 3x_3 - 2x_4 + 2x_5 &= 5,0;\\ x_2 - 2x_3 + 5x_4 + 3x_5 &= 7,5;\\ -2x_1 - 3x_2 + 2x_3 + 3x_4 + 4x_5 &= 3,3. \end{aligned}\right\}$$

Solution. Inscrivons les coefficients a_{ij} et les termes constants b_i du système considéré dans la section initiale A du tableau

Tableau 18

Résolution d'un système linéaire par la méthode des racines carrées

a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}	b_i	Σ	Sections du schéma
1	3	−2	0	−2	0,5	0,5	*A*
3	4	−5	1	−3	5,4	5,4	
−2	−5	3	−2	2	5,0	1,0	
0	1	−2	5	3	7,5	14,5	
−2	−3	2	3	4	3,3	7,3	
t_{i1}	t_{i2}	t_{i3}	t_{i4}	t_{i5}	y_i	Σ	
1	3	−2	0	−2	0,5	0,5	*B*
	2,2361i	−0,4472i	−0,4472i	−1,3416i	−1,7471i	−1,7471i	
		0,8944i	2,0125i	1,5653i	−7,5803i	−3,1081i	
			3,0414	2,2194	−2,2928	2,9679	
				0,8221i	0,1643i	0,9859i	
−6,0978	−2,2016	−6,8011	−0,8996	0,1998		$\bar{x}_i$	*C*
−5,0973	−1,2017	−5,8004	0,1007	1,1992		x_i	

(tableau 18) et calculons la colonne Σ. En appliquant les formules (3) et (6) et en passant successivement d'une ligne à l'autre, calculons les coefficients t_{ij} et les nouveaux termes constants y_i pour compléter de cette façon la section B du tableau.

Par exemple,

$$t_{35}=\frac{a_{35}-t_{13}t_{15}-t_{23}t_{25}}{t_{33}}=\frac{2-(-2)(-2)-(-0{,}4472i)(-1{,}3416i)}{0{,}8944i}=1{,}5653i.$$

Pour vérifier calculons la colonne Σ. Calculons d'après les formules (7) les valeurs des inconnues x_i et les grandeurs de contrôle $\bar{x}_i = x_i + 1$, en les portant sur la section C. Par exemple,

$$x_3=\frac{y_3-t_{35}x_5-t_{34}x_4}{t_{33}}=$$
$$=\frac{-7{,}5803i-1{,}5652i\cdot 0{,}1998-2{,}0125i\cdot(-0{,}8996)}{0{,}8944i}=-6{,}8011.$$

§ 9. Schéma de Khaletski

Pour la commodité du raisonnement écrivons le système d'équations linéaires sous une forme matricielle

$$Ax = b, \tag{1}$$

où $A = [a_{ij}]$ est une matrice carrée d'ordre n et

$$x=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad b=\begin{bmatrix} a_{1,\,n+1} \\ \cdots \\ a_{n,\,n+1} \end{bmatrix}$$

sont des vecteurs colonnes. Mettons la matrice A sous forme d'un produit de la matrice triangulaire inférieure $B = [b_{ij}]$ et de la matrice triangulaire supérieure $C = [c_{ij}]$ à diagonale unité, c'est-à-dire

$$A = BC, \tag{2}$$

où

$$B=\begin{bmatrix} b_{11} & 0 & \dots & 0 \\ b_{21} & b_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix} \quad \text{et} \quad C=\begin{bmatrix} 1 & c_{12} & \dots & c_{1n} \\ 0 & 1 & \dots & c_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Les éléments b_{ij} et c_{ij} se définissent alors d'après les formules

$$\left.\begin{aligned} &b_{i1}=a_{i1}, \\ &b_{ij}=a_{ij}-\sum_{k=1}^{j-1} b_{ik}c_{kj} \quad (i \geqslant j > 1) \end{aligned}\right\} \tag{3}$$

et

$$\left.\begin{aligned} c_{1j} &= \frac{a_{1j}}{b_{11}}, \\ c_{ij} &= \frac{1}{b_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} b_{ik}c_{kj}\right) \quad (1 < i < j). \end{aligned}\right\} \tag{4}$$

D'où le vecteur recherché x peut être calculé d'après la chaîne d'équations

$$By = b, \quad Cx = y. \tag{5}$$

Les matrices B et C étant triangulaires, les systèmes (5) se résolvent sans peine

$$\left.\begin{aligned} y_1 &= \frac{a_{1,\, n+1}}{b_{11}}, \\ y_i &= \frac{1}{b_{ii}}\left(a_{i,\, n+1} - \sum_{k=1}^{i-1} b_{ik}y_k\right) \quad (i > 1) \end{aligned}\right\} \tag{6}$$

et

$$\left.\begin{aligned} x_n &= y_n, \\ x_i &= y_i - \sum_{k=i+1}^{n} c_{ik}x_k \quad (i < n). \end{aligned}\right\} \tag{7}$$

Les formules (6) montrent qu'il est avantageux de calculer les nombres y_i simultanément avec les coefficients c_{ij}. Cette méthode a reçu le nom de *schéma de Khaletski*. Ce schéma fait appel au contrôle usuel à l'aide des sommes.

Remarquons que si la matrice A est symétrique, c'est-à-dire si $a_{ij}=a_{ji}$,

$$c_{ij} = \frac{b_{ji}}{b_{ii}} \quad (i < j).$$

Le schéma de Khaletski est commode pour travailler sur des calculatrices électroniques, car dans ce cas les opérations de « mémorisation » (3) et (4) peuvent se faire sans enregistrer les résultats intermédiaires.

Exemple. Résoudre le système

$$\left.\begin{aligned} 3x_1 + x_2 - x_3 + 2x_4 &= 6; \\ -5x_1 + x_2 + 3x_3 - 4x_4 &= -12; \\ 2x_1 + x_3 - x_4 &= 1; \\ x_1 - 5x_2 + 3x_3 - 3x_4 &= 3. \end{aligned}\right\}$$

Tableau 19

	x_1	x_2	x_3	x_4		Σ	x_1	x_2	x_3	x_4		Σ
I	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	3	1	−1	2	6	11
	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	−5	1	3	−4	−12	−17
	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	2	0	1	−1	1	3
	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	1	−5	3	−3	3	−1
II	b_{11} 1	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	3 1	0,333333	−0,333333	0,666667	2	3,666667
	b_{21}	b_{22} 1	c_{23}	c_{24}	c_{25}	c_{26}	−5	2,666667 1	0,5	−0,25	−0,75	0,5
	b_{31}	b_{32}	b_{33} 1	c_{34}	c_{35}	c_{36}	2	−0,666667	2 1	−1,25	−1,75	−2
	b_{41}	b_{42}	b_{43}	b_{44} 1	c_{45}	c_{46}	1	−5,333333	6	2,5 1	3	4
III					y_1	x_1					2	1
					y_2	x_2					−0,75	−1
					y_3	x_3					−1,75	2
					y_4	x_4					3	3

S o l u t i o n. (Cf. tableau 19).

Ecrivons dans la première section du tableau 19 la matrice des coefficients du système, ses termes constants et les sommes de contrôle.

Ensuite, puisque $b_{i1} = a_{i1}$ ($i = 1, 2, 3, 4$), la première colonne de la section I est reportée dans la première colonne de la section II.

Pour obtenir la première ligne de la section II, divisons tous les éléments de la première ligne de la section I par l'élément $a_{11} = b_{11}$, dans notre cas par 3.

On a:

$$c_{12} = \frac{1}{3} = 0,(3);$$

$$c_{13} = -\frac{1}{3} = -0,(3);$$

$$c_{14} = \frac{2}{3} = 0,(6);$$

$$c_{15} = \frac{6}{3} = 2;$$

$$c_{16} = \frac{11}{3} = 2,(6).$$

Complétons maintenant la deuxième colonne de la section II en partant de la deuxième ligne. En appliquant les formules (3) nous déterminons b_{j2}:

$$b_{22} = a_{22} - b_{21}c_{12} = 1 - \left(-5 \cdot \frac{1}{3}\right) = \frac{8}{3} = 2,66\,(6);$$

$$b_{32} = a_{32} - b_{31}c_{12} = 0 - 2 \cdot \frac{1}{3} = -\frac{2}{3} = 0,(6);$$

$$b_{42} = a_{42} - b_{41}c_{12} = -5 - 1 \cdot \frac{1}{3} = -5\frac{1}{3} = -5,(3).$$

Ensuite, le calcul de c_{2j} ($j = 3, 4, 5, 6$) d'après les formules (4) permet de composer la deuxième ligne de la section II:

$$c_{23} = \frac{1}{b_{22}}(a_{23} - b_{21}c_{13}) = \frac{3}{8}\left[3 - (-5) \cdot \left(-\frac{1}{3}\right)\right] = \frac{1}{2};$$

$$c_{24} = \frac{1}{b_{22}}(a_{24} - b_{21}c_{14}) = \frac{3}{8}\left[(-4) - (-5) \cdot \frac{2}{3}\right] = -\frac{1}{4};$$

$$c_{25} = \frac{1}{b_{22}}(a_{25} - b_{21}c_{15}) = \frac{3}{8}\left[(-12) - (-5) \cdot 2\right] = -\frac{3}{4};$$

$$c_{26} = \frac{1}{b_{22}}(a_{26} - b_{21}c_{16}) = \frac{3}{8}\left[(-17) - (-5) \cdot \frac{11}{3}\right] = \frac{1}{2}.$$

Passons à la troisième colonne et calculons ses éléments b_{33} et b_{34} d'après les formules (3), etc., tant qu'on ne complète toute la sec-

tion II. Ainsi, la section II se forme en écrivant tour à tour colonne — ligne, colonne — ligne, etc.

En se servant des formules (6) et (7), on détermine dans la section III y_i et x_i $(i = 1, 2, 3, 4)$.

La vérification courante s'effectue à l'aide de la colonne Σ qui subit les mêmes opérations que la colonne des termes constants.

§ 10. Méthode des approximations successives

Lorsque le nombre des inconnues d'un système linéaire est grand, le schéma de la méthode de Gauss qui donne une solution exacte devient trop compliqué. Dans ces conditions il devient beaucoup plus commode de rechercher les racines du système par des méthodes numériques approchées. L'une d'elles, la méthode des *approximations successives* dite aussi *méthode des itérations*, est exposée dans ce qui suit.

Soit le système linéaire

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{array}\right\} \qquad (1)$$

Introduisons dans la discussion les matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

et mettons le système (1) sous forme d'une équation matricielle

$$Ax = b. \qquad (1')$$

En supposant que les coefficients diagonaux

$$a_{ii} \neq 0 \qquad (i = 1, 2, \ldots, n),$$

on résout la première équation du système (1) par rapport à x_1, la deuxième par rapport à x_2, etc. On a alors le système équivalent

$$\left.\begin{array}{l} x_1 = \beta_1 + \alpha_{12}x_2 + \alpha_{13}x_3 + \ldots + \alpha_{1n}x_n, \\ x_2 = \beta_2 + \alpha_{21}x_1 + \alpha_{23}x_3 + \ldots + \alpha_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = \beta_n + \alpha_{n1}x_1 + \alpha_{n2}x_2 + \ldots + \alpha_{n,\,n-1}x_{n-1}, \end{array}\right\} \qquad (2)$$

où

$$\beta_i = \frac{b_i}{a_{ii}}; \quad \alpha_{ij} = -\frac{a_{ij}}{a_{ii}} \qquad \text{pour } i \neq j$$

et $a_{ij} = 0$ pour $i = j$ $(i, j = 1, 2, \ldots, n)$.

Introduisons les matrices

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{bmatrix} \quad \text{et} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix},$$

pour mettre le système (2) sous une forme matricielle

$$x = \beta + \alpha x. \tag{2'}$$

Cherchons la solution du système (2) par la *méthode des approximations successives*. Prenons, par exemple, pour approximation initiale la colonne des termes constants $x^{(0)} = \beta$.

Puis construisons successivement les matrices colonnes

$$x^{(1)} = \beta + \alpha x^{(0)}$$

(*première approximation*),

$$x^{(2)} = \beta + \alpha x^{(1)}$$

(*deuxième approximation*), etc.

Toute $(k+1)^{\text{ième}}$ approximation se calcule en général d'après la formule

$$x^{(k+1)} = \beta + \alpha x^{(k)} \qquad (k = 0, 1, 2, \ldots). \tag{3}$$

Si la suite des approximations $x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots$ possède une limite

$$x = \lim_{k \to \infty} x^{(k)},$$

cette limite est une solution du système (2). En effet, en passant à la limite dans l'égalité (3), on a:

$$\lim_{k \to \infty} x^{(k+1)} = \beta + \alpha \lim_{k \to \infty} x^{(k)}$$

ou

$$x = \beta + \alpha x,$$

c'est-à-dire le vecteur limite x est une solution du système (2') et par conséquent du système (1).

Ecrivons les formules des approximations sous une forme développée:

$$\left.\begin{aligned} &x_i^{(0)} = \beta_i, \\ &x_i^{(k+1)} = \beta_i + \sum_{j=1}^{n} \alpha_{ij} x_j^{(k)} \\ &(\alpha_{ii} = 0;\ i = 1, \ldots, n;\ k = 0, 1, 2, \ldots). \end{aligned}\right\} \qquad (3')$$

Remarquons que parfois il est plus avantageux de ramener le système (1) au type (2) de façon à ne pas annuler les coefficients α_{ii}. Il serait logique, par exemple, pour appliquer la méthode des approximations successives, d'écrire l'équation

$$1{,}02x_1 - 0{,}15x_2 = 2{,}7$$

sous la forme

$$x_1 = 2{,}7 - 0{,}02x_1 + 0{,}15x_2.$$

En général, dans le cas du système

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad (i = 1, 2, \ldots, n),$$

on peut poser:

$$a_{ii} = a_{ii}^{(1)} + a_{ii}^{(2)},$$

avec $a_{ii}^{(1)} \neq 0$. Le système en question est alors équivalent au système réduit

$$x_i = \beta_i + \sum_{j=1}^{n} \alpha_{ij} x_j \qquad (i = 1, 2, \ldots, n),$$

où

$$\beta_i = \frac{b_i}{a_{ii}^{(1)}}, \quad \alpha_{ii} = -\frac{a_{ii}^{(2)}}{a_{ii}^{(1)}}, \quad \alpha_{ij} = -\frac{a_{ij}}{a_{ii}^{(1)}} \quad \text{pour } i \neq j.$$

C'est pourquoi, dans nos raisonnements nous ne supposons généralement pas que $\alpha_{ii} = 0$.

La méthode des approximations successives définies par les formules (3) ou (3′) s'appelle également *méthode itérative*. Le processus itératif (3) converge bien, c'est-à-dire le nombre d'approximations nécessaires pour obtenir les solutions du système (1) avec la précision imposée n'est pas grand si les éléments de la matrice α sont petits en valeur absolue. Autrement dit, pour appliquer avec succès le processus itératif, il faut que les modules des coefficients diagonaux du système (1) soient grands par rapport aux modules des coefficients non diagonaux de ce système (les termes constants ne jouent alors aucun rôle).

Exemple 1. Résoudre le système

$$\left.\begin{aligned} 4x_1+0{,}24x_2-0{,}08x_3&=8,\\ 0{,}09x_1+3x_2-0{,}15x_3&=9,\\ 0{,}04x_1-0{,}08x_2+4x_3&=20\end{aligned}\right\}\qquad(4)$$

par la méthode des approximations successives.

Solution. Les coefficients diagonaux 4; 3; 4 du système dominent nettement ici les autres coefficients des inconnues. En ramenant ce système à la forme normale (2), on a

$$\left.\begin{aligned} x_1&=2-0{,}06x_2+0{,}02x_3,\\ x_2&=3-0{,}03x_1+0{,}05x_3,\\ x_3&=5-0{,}01x_1+0{,}02x_2.\end{aligned}\right\}\qquad(5)$$

Le système (5) peut s'écrire sous la forme matricielle:

$$\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}2\\3\\5\end{bmatrix}+\begin{bmatrix}0&-0{,}06&0{,}02\\-0{,}03&0&0{,}05\\-0{,}01&0{,}02&0\end{bmatrix}\cdot\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}.$$

Pour les approximations initiales de la solution de (4) on prend:

$$x_1^{(0)}=2;\quad x_2^{(0)}=3;\quad x_3^{(0)}=5.$$

En portant ces valeurs dans les seconds membres de (5), on obtient les premières approximations de la solution

$$x_1^{(1)}=2-0{,}06\cdot3+0{,}02\cdot5=1{,}92;$$
$$x_2^{(1)}=3-0{,}03\cdot2+0{,}05\cdot5=3{,}19;$$
$$x_3^{(1)}=5-0{,}01\cdot2+0{,}02\cdot3=5{,}04.$$

Ensuite, en portant ces approximations trouvées dans la formule (5), on obtient les deuxièmes approximations de la solution:

$$x_1^{(2)}=1{,}9094;\quad x_2^{(2)}=3{,}1944;\quad x_3^{(2)}=5{,}0446.$$

Après une nouvelle substitution, on obtient les troisièmes approximations des solutions

$$x_1^{(3)}=1{,}90923;\quad x_2^{(3)}=3{,}19495;\quad x_3^{(3)}=5{,}04485,\ \text{etc.}$$

Les résultats des calculs sont portés sur le tableau 20.

Remarque. Lors de l'application de la méthode des approximations successives (formule (3)) aucun besoin n'est de prendre la colonne des termes constants pour approximation initiale $x^{(0)}$. Nous allons montrer dans ce qui suit que la convergence du processus itératif ne dépend que des propriétés de la matrice α; de plus, certaines conditions étant observées, si ce processus converge pour un choix quelconque de l'approximation initiale, il convergera également vers le même vecteur limite pour tout autre choix de cette

Tableau 20

Résolution du système linéaire par la méthode des approximations successives

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	2	2	5
1	1,92	3,19	5,04
2	1,9094	3,1944	5,0446
3	1,90923	3,19495	5,04485

approximation. C'est pourquoi dans le processus itératif le vecteur initial $x^{(0)}$ peut être arbitraire. Il est raisonnable de prendre comme vecteur initial une solution approchée du système établie par approximation grossière.

Le processus itératif convergent jouit de la propriété importante d'autocorrection, qui fait qu'une erreur de calcul isolée n'entache pas le résultat final, une approximation erronée pouvant être considérée comme un nouveau vecteur initial.

Constatons qu'il est parfois plus commode de calculer non pas les approximations elles-mêmes, mais leurs différences. En introduisant les notations

$$\Delta^{(k)} = x^{(k)} - x^{(k-1)} \qquad (k = 0, 1, 2, \ldots),$$

on obtient d'après la formule (3):

$$x^{(k+1)} = \beta + \alpha x^{(k)} \tag{6}$$

et

$$x^{(k)} = \beta + \alpha x^{(k-1)}. \tag{7}$$

D'où, en retranchant l'égalité (7) de l'égalité (6),

$$\Delta^{(k+1)} = \alpha\,(x^{(k)} x - ^{(k-1)}) = \alpha\Delta^{(k)},$$

soit

$$\Delta^{(k+1)} = \alpha\Delta^{(k)} \qquad (k = 1, 2, \ldots). \tag{8}$$

On prend pour approximation initiale

$$\Delta^{(0)} = x^{(0)}, \tag{9}$$

alors la m-ième approximation s'écrit

$$x^{(m)} = \sum_{k=0}^{m} \Delta^{(k)}. \tag{10}$$

Si l'on adopte comme d'ordinaire $\Delta^{(0)} = x^{(0)} = \beta$, l'égalité (8) est vérifiée également avec $k = 0$. Dans le cas contraire, pour $k = 0$ l'égalité (8) n'a pas lieu. On en déduit la méthode suivante de calcul de cette variante de l'itération:

1) si $\Delta^{(0)} = x^{(0)} = \beta$, il vient

$$\Delta^{(k)} = \alpha\Delta^{(k-1)} = \alpha^k\beta \quad (k = 0,\ 1,\ 2,\ \ldots)$$

et

$$x^{(k)} = \sum_{s=0}^{k} \Delta^{(s)} = \sum_{s=0}^{k} \alpha^s\beta\,;$$

2) si $\Delta^{(0)} = x^{(0)} \neq \beta$, on trouve

$$\Delta^{(1)} = x^{(1)} - x^{(0)} = \alpha x^{(0)} + \beta - x^{(0)}$$

pour poser

$$\Delta^{(k)} = \alpha\Delta^{(k-1)} = \alpha^{k-1}\Delta^{(1)} \quad (k = 1,\ 2,\ 3,\ \ldots)$$

Par conséquent

$$x^{(k)} = \sum_{s=0}^{k} \Delta^{(s)} = x^{(0)} + \sum_{s=1}^{k} \alpha^{s-1}\Delta^{(1)}.$$

Exemple 2. Résoudre le système

$$\left.\begin{aligned} 2x_1 - x_2 + x_3 &= -3, \\ 3x_1 + 5x_2 - 2x_3 &= 1, \\ x_1 - 4x_2 + 10x_3 &= 0. \end{aligned}\right\} \qquad (11)$$

Solution. Ramenons le système (11) à la forme (2):

$$\begin{aligned} x_1 &= -1{,}5 + 0{,}5x_2 - 0{,}5x_3\,; \\ x_2 &= 0{,}2 - 0{,}6x_1 + 0{,}4x_3\,; \\ x_3 &= -0{,}1x_1 + 0{,}4x_2. \end{aligned}$$

Ici

$$\alpha = \begin{bmatrix} 0 & 0{,}5 & -0{,}5 \\ -0{,}6 & 0 & 0{,}4 \\ -0{,}1 & 0{,}4 & 0 \end{bmatrix}$$

et

$$\beta = \begin{bmatrix} -1{,}5 \\ 0{,}2 \\ 0 \end{bmatrix}.$$

Les formules (8) et (9) donnent :

$$\Delta^{(0)} = \beta = \begin{bmatrix} -1{,}5 \\ 0{,}2 \\ 0 \end{bmatrix};$$

$$\Delta^{(1)} = \alpha\Delta^{(0)} = \begin{bmatrix} 0 & 0{,}5 & -0{,}5 \\ -0{,}6 & 0 & 0{,}4 \\ -0{,}1 & 0{,}4 & 0 \end{bmatrix} \begin{bmatrix} -1{,}5 \\ 0{,}2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}1 \\ 0{,}9 \\ 0{,}23 \end{bmatrix};$$

$$\Delta^{(2)} = \alpha\Delta^{(1)} = \begin{bmatrix} 0 & 0{,}5 & -0{,}5 \\ -0{,}6 & 0 & 0{,}4 \\ -0{,}1 & 0{,}4 & 0 \end{bmatrix} \begin{bmatrix} 0{,}1 \\ 0{,}9 \\ 0{,}23 \end{bmatrix} = \begin{bmatrix} 0{,}335 \\ 0{,}032 \\ 0{,}350 \end{bmatrix},$$

etc. Les résultats sont portés sur le tableau 21.

Tableau 21

Résolution du système linéaire par la méthode modifiée des approximations successives

(méthode cumulative)

k	$\Delta_{x_1}^{(k)}$	$\Delta_{x_2}^{(k)}$	$\Delta_{x_3}^{(k)}$
0	−1,500	0,200	0,000
1	0,100	0,900	0,230
2	0,335	0,032	0,350
3	−0,159	−0,061	−0,021
4	−0,020	0,011	−0,008
5	0,010	0,009	0,006
6	0,002	−0,004	0,003
7	−0,004	0,000	−0,001
8	0,000	0,002	0,000
9	0,001	0,000	0,001
Σ	−1,235	1,089	0,560

Ainsi, les solutions approchées sont :

$$x_1 = -1{,}235; \quad x_2 = 1{,}089; \quad x_3 = 0{,}560.$$

L'inconvénient de cette variante de la méthode des approximations successives est le cumul systématique des erreurs avec l'augmentation du nombre de termes, ce qui conduit éventuellement à des erreurs importantes dans les solutions cherchées. De plus, l'erreur de calcul influe sur le résultat final. La première variante de la méthode des approximations successives est donc plus sûre.

Remarques sur la précision de calcul. Si tous les coefficients et termes constants du système donné sont des nombres exacts, sa résolution par la méthode des approximations successives peut s'obtenir avec n'importe quel nombre m de chiffres exacts, donné à l'avance. Dans les valeurs des approximations successives il faut retenir $m+1$ chiffres et calculer les approximations successives tant qu'elles ne se confondent, après quoi le résultat est arrondi d'un chiffre. Si les coefficients et les termes constants du système considéré sont des nombres approchés écrits avec p chiffres, la résolution de ce système se fait de même que dans le cas des nombres exacts avec $m = p$ chiffres significatifs.

Voici sans démonstration une condition suffisante de la convergence du processus itératif (la démonstration est donnée au chapitre IX, § 1).

Théorème. *Si au moins une des conditions*

$$1) \qquad \sum_{j=1}^{n} |\alpha_{ij}| < 1 \quad (i = 1, 2, \ldots, n)$$

ou

$$2) \qquad \sum_{i=1}^{n} |\alpha_{ij}| < 1 \quad (j = 1, 2, \ldots, n)$$

est vraie pour le système réduit (2), *le processus itératif* (3) *converge vers la solution unique de ce système, quel que soit le choix de l'approximation initiale.*

Corollaire. Pour le système

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad (i = 1, 2, \ldots, n)$$

la méthode des approximations successives converge si les inégalités

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad (i = 1, 2, \ldots, n)$$

sont vérifiées, c'est-à-dire si pour toute équation du système le coefficient diagonal est plus grand en module que la somme des modules de tous les autres coefficients (sans termes constants).

§ 11. Réduction d'un système linéaire à la forme commode pour l'itération

Le théorème de la convergence (§ 10) impose des conditions serrées aux coefficients du système linéaire considéré

$$Ax = b. \qquad (1)$$

Toutefois, si $\det A \neq 0$, la combinaison linéaire des équations du système (15) permet toujours de remplacer ce dernier par un système équivalent

$$x = \beta + \alpha x, \tag{2}$$

tel que les conditions du théorème de convergence seront remplies.

En effet, multiplions l'équation (1) par la matrice $D = A^{-1} - \varepsilon$, où $\varepsilon = [\varepsilon_{ij}]$ est une matrice à éléments petits en module. On a alors:

$$(A^{-1} - \varepsilon) Ax = Db$$

ou

$$x = \beta + \alpha x, \tag{3}$$

avec $\alpha = \varepsilon A$ et $\beta = Db$. Si $|\varepsilon_{ij}|$ sont suffisamment petits, il est clair que le système (3) vérifie les conditions du théorème de convergence.

La multiplication par la matrice D est équivalente à l'ensemble des transformations élémentaires sur les équations du système. La tâche consiste à aboutir au type standard (3) avec le moins d'efforts possible.

En pratique on procède de la façon suivante. Dans le système concerné on prélève les équations dont les coefficients sont supérieurs en module à la somme des modules des autres coefficients de l'équation. Toute équation prélevée s'inscrit sur une ligne du nouveau système de façon que le coefficient le plus grand en module soit diagonal.

Les équations restantes et les équations prélevées du système sont associées en combinaisons linéaires linéairement indépendantes de sorte que soit observé le principe de complétation du nouveau système décrit ci-dessus et que toutes les lignes vides soient remplies. Il faut de plus prendre soin pour que toute équation non utilisée fasse partie au moins d'une combinaison linéaire qui constitue une équation du nouveau système. Donnons un exemple pour illustrer tout ce qui vient d'être dit.

Exemple. Ramener le système

$$\left.\begin{array}{ll}(A) & 2x_1+3x_2-4x_3+x_4-3=0,\\ (B) & x_1-2x_2-5x_3+x_4-2=0,\\ (C) & 5x_1-3x_2+x_3-4x_4-1=0,\\ (D) & 10x_1+2x_2-x_3+2x_4+4=0\end{array}\right\}$$

au type commode pour l'itération.

Solution. Le coefficient de x_3 de l'équation (B) étant plus grand en module que la somme des modules des autres coefficients, on peut considérer cette équation comme la troisième équation d'un nouveau système. Dans l'équation (D) le coefficient de x_1 est égale-

ment plus grand que la somme des modules des autres coefficients; cette équation peut donc être prise pour la première équation du nouveau système. Par conséquent le nouveau système s'écrit:

$$\left.\begin{array}{ll}(\mathrm{I}) & 10x_1+2x_2-\ x_3+2x_4+4=0,\\ (\mathrm{II}) & \ldots\ldots\ldots\ldots\\ (\mathrm{III}) & x_1-2x_2-5x_3+\ x_4-2=0,\\ (\mathrm{IV}) & \ldots\ldots\ldots\ldots\end{array}\right\}$$

En analysant le système donné on voit sans peine que pour obtenir l'équation (II) au coefficient de x_2 maximal en module, il suffit de composer la différence $(A)-(C)$:

$$(\mathrm{II})\qquad x_1+5x_2+x_3+0x_4-1=0.$$

Maintenant le nouveau système comprend les équations (A), (B) et (D); par suite l'équation (IV) contient nécessairement l'équation (C) du système donné. La sélection montre que pour l'équation (IV) on peut prendre la combinaison linéaire $2(A)-(B)-2(C)-(D)$, c'est-à-dire

$$(\mathrm{IV})\qquad 3x_1+0x_2+0x_3-9x_4-10=0.$$

Finalement on obtient le système transformé d'équations I-IV équivalent au système initial et vérifiant les conditions de convergence du processus itératif. La résolution de ce système par rapport aux inconnues diagonales conduit au système

$$\left.\begin{array}{l}x_1=0x_1\quad -0{,}2x_2+0{,}1x_3-0{,}2x_4-0{,}4;\\ x_2=0{,}2x_1\quad +0x_2\quad -0{,}2x_3+0x_4\quad +0{,}2;\\ x_3=0{,}2x_1\quad -0{,}4x_2+0x_3\quad +0{,}2x_4-0{,}4;\\ x_4=0{,}333x_1+0x_2\quad +0x_3\quad +0x_4\quad -1{,}111,\end{array}\right\}$$

auquel on peut appliquer la méthode des approximations successives.

§ 12. Méthode de Seidel

La méthode de Seidel est une modification de la méthode des approximations successives. Son idée maîtresse consiste à tenir compte, lors du calcul de la $(k+1)$-ième approximation de l'inconnue x_i, des $(k+1)$-ièmes approximations des inconnues x_1, x_2, . . ., x_{i-1} déjà établies.

Soit le système linéaire réduit

$$x_i=\beta_i+\sum_{j=1}^{n}\alpha_{ij}x_j\quad (i=1,\ 2,\ \ldots,\ n).$$

Choisissons arbitrairement les approximations initiales de la solution

$$x_1^{(0)},\ x_2^{(0)},\ \ldots,\ x_n^{(0)},$$

en prenant garde, certes, qu'elle correspond dans une certaine mesure aux inconnues cherchées

$$x_1,\ x_2,\ \ldots,\ x_n.$$

Ensuite, en supposant que les k-ièmes approximations $x_i^{(k)}$ $(i = 1, \ldots, n)$ sont connues, nous construisons suivant Seidel les $(k + 1)$-ièmes approximations de la solution d'après les formules suivantes:

$$x_1^{(k+1)} = \beta_1 + \sum_{j=1}^{n} \alpha_{1j} x_j^{(k)};$$

$$x_2^{(k+1)} = \beta_2 + \alpha_{21} x_1^{(k+1)} + \sum_{j=2}^{n} \alpha_{2j} x_j^{(k)};$$

. . . .

$$x_i^{(k+1)} = \beta_i + \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k)};$$

. . . .

$$x_n^{(k+1)} = \beta_n + \sum_{j=1}^{n-1} \alpha_{nj} x_j^{(k+1)} + \alpha_{nn} x_n^{(k)} \quad (k = 0,\ 1,\ 2,\ \ldots).$$

Notons que le théorème de convergence pour une itération simple donné au § 10 reste vrai pour une itération suivant la méthode de Seidel (cf. chapitre IX, §§ 3 à 7).

Dans les cas courants la méthode de Seidel assure une meilleure convergence que la méthode itérative simple, mais généralement son application est plus délicate. Le processus de Seidel peut converger même dans le cas où le processus itératif simple diverge. Cependant il n'en est pas toujours ainsi. Il arrive que le processus de Seidel converge plus lentement que le processus itératif simple. Bien plus, dans certains cas le processus itératif converge alors que le processus de Seidel est divergent [1] (cf. chapitre XI, § 6).

E x e m p l e. Résoudre par la méthode de Seidel le système d'équations

$$\left.\begin{aligned} 10x_1 + x_2 + x_3 &= 12; \\ 2x_1 + 10x_2 + x_3 &= 13; \\ 2x_1 + 2x_2 + 10x_3 &= 14. \end{aligned}\right\}$$

S o l u t i o n. Ramenons ce système à la forme commode pour l'itération

$$\left.\begin{aligned} x_1 &= 1{,}2-0{,}1x_2-0{,}1x_3\,; \\ x_2 &= 1{,}3-0{,}2x_1-0{,}1x_3\,; \\ x_3 &= 1{,}4-0{,}2x_1-0{,}2x_2. \end{aligned}\right\}$$

Prenons pour approximations initiales

$$x_1^{(0)}=1{,}2\,; \quad x_2^{(0)}=0\,; \quad x_3^{(0)}=0.$$

En appliquant successivement le processus de Seidel, on a:

$$\left.\begin{aligned} x_1^{(1)} &= 1{,}2-0{,}1\cdot 0 \quad -0{,}1\cdot 0 \quad\;\; =1{,}2\,; \\ x_2^{(1)} &= 1{,}3-0{,}2\cdot 1{,}2-0{,}1\cdot 0 \quad\;\; =1{,}06\,; \\ x_3^{(1)} &= 1{,}4-0{,}2\cdot 1{,}2-0{,}2\cdot 1{,}06=0{,}948\,; \end{aligned}\right\} \qquad \text{(I)}$$

$$\left.\begin{aligned} x_1^{(2)} &= 1{,}2-0{,}1\cdot 1{,}06-0{,}1\cdot 0{,}948=0{,}9992\,; \\ x_2^{(2)} &= 1{,}3-0{,}2\cdot 0{,}9992-0{,}1\cdot 0{,}948=1{,}00536\,; \\ x_3^{(2)} &= 1{,}4-0{,}2\cdot 0{,}9992-0{,}2\cdot 1{,}00536=0{,}999098, \text{ etc.} \end{aligned}\right\} \qquad \text{(II)}$$

Les résultats du calcul avec quatre décimales exactes sont portés sur le tableau 22.

Tableau 22

Recherche des racines d'un système linéaire par la méthode de Seidel

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	1,2000	0,0000	0,0000
1	1,2000	1,0600	0,9480
2	0,9992	1,0054	0,9991
3	0,9996	1,0001	1,0001
4	1,0000	1,0000	1,0000
5	1,0000	1,0000	1,0000

La solution exacte est: $x_1 = 1$; $x_2 = 1$; $x_3 = 1$.

§ 13. Cas d'un système normal

D é f i n i t i o n 1. Un polynôme entier homogène du second degré de n variables s'appelle *forme quadratique* de ces variables. Dans le cas général, la forme quadratique s'écrit

$$u\,(x_1, x_2, \ldots, x_n) = a_{11}x_1^2 + a_{22}x_2^2 + \ldots + a_{nn}x_n^2 + \\ + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \ldots + 2a_{n-1,n}x_{n-1}x_n, \qquad (1)$$

où a_{ij} $(i, j = 1, 2, \ldots, n)$ sont des nombres constants; pour $i \neq j$, par souci de commodité, les coefficients respectifs sont pris pairs: $2a_{ij}$. En égalant u à la constante c, on obtient l'équation d'une quadrique à centre

$$u(x_1, x_2, \ldots, x_n) = c$$

dans un espace de dimension n.

Si l'on pose

$$a_{ij} = a_{ji}, \tag{2}$$

c'est-à-dire $2a_{ij} = a_{ij} + a_{ji}$, l'écriture de la formule (1) peut être abrégée:

$$u(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j. \tag{1'}$$

La matrice

$$A = [a_{ij}] \tag{3}$$

s'appelle *matrice de la forme quadratique* (1'). En vertu de la condition (2) la matrice A est symétrique, c'est-à-dire elle coïncide avec sa transposée. Inversement, pour toute matrice symétrique $A = [a_{ij}]$ on peut construire une forme quadratique correspondante (1').

D é f i n i t i o n 2. Une forme quadratique (1) est dite *définie positive* (*négative*) si elle prend des valeurs positives (négatives) et ne s'annule qu'avec

$$x_1 = x_2 = \ldots = x_n = 0.$$

Si $u(x_1, x_2, \ldots, x_n)$ est une forme quadratique définie positive, l'équation

$$u(x_1, x_2, \ldots, x_n) = c \qquad (c > 0)$$

est l'équation d'ellipsoïde. Notons que dans ce cas

$$a_{ii} > 0 \qquad (i = 1, 2, \ldots, n),$$

du fait que

$$\begin{aligned} a_{11} &= u(1, 0, \ldots, 0) > 0, \\ a_{22} &= u(0, 1, \ldots, 0) > 0, \\ &\ldots\ldots\ldots\ldots \\ a_{nn} &= u(0, 0, \ldots, 1) > 0. \end{aligned}$$

D é f i n i t i o n 3. Le système linéaire

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad (i = 1, 2, \ldots, n) \tag{4}$$

est dit *normal* si: 1) la matrice des coefficients $A = [a_{ij}]$ est symétrique, c'est-à-dire si $a_{ij} = a_{ji}$, 2) la forme quadratique correspon-

dante

$$u = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

est définie positive.

Les systèmes normaux se présentent dans la résolution de nombreux problèmes, et, entre autres, dans la méthode des moindres carrés, la recherche de la direction des axes principaux d'un ellipsoïde, etc.

Ramenons le système normal (4) par le procédé usuel à la forme spéciale

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \alpha_{ij} x_j + \beta_i, \tag{4'}$$

où

$$\alpha_{ij} = -\frac{a_{ij}}{a_{ii}} \quad (j \neq i) \text{ et } \beta_i = \frac{b_i}{a_{ii}}.$$

Théorème 1. *Si le système linéaire* (4) *est un système normal, le processus de Seidel du système réduit* (4') *équivalent est toujours convergent.*

Démonstration cf. chapitre XI, § 5, ainsi que [2].

Le mode de réduction d'un système linéaire à la forme normale est décrit par le théorème qui suit.

Théorème 2. *Si les deux membres d'un système linéaire*

$$Ax = b \tag{5}$$

à matrice régulière $A = [a_{ij}]$ *sont multipliés à gauche par la transposée* $A' = [a_{ji}]$, *le nouveau système*

$$A'Ax = A'b \tag{6}$$

sera un système normal.

Démonstration. Montrons d'abord que la matrice $A'A$ est symétrique. En effet, on a:

$$(A'A)' = A'A'' = A'A.$$

Montrons maintenant que la forme quadratique associée à la matrice $A'A$ est définie positive. Composons la forme quadratique à matrice $A'A$:

$$u(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ki} a_{kj} x_i x_j.$$

En changeant l'ordre de sommation, on obtient:

$$u = \sum_{k=1}^{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ki} x_i a_{kj} x_j = \sum_{k=1}^{n} \left(\sum_{i=1}^{n} a_{ki} x_i \cdot \sum_{j=1}^{n} a_{kj} x_j \right).$$

La valeur de la somme ne dépendant pas de la notation de l'indice de sommation, on a:

$$u = \sum_{k=1}^{n} \left(\sum_{i=1}^{n} a_{ki} x_i \right)^2 \geqslant 0.$$

Par hypothèse, $\det A = \det [a_{ij}] \neq 0$. Donc le système homogène

$$\sum_{i=1}^{n} a_{ki} x_i = 0 \qquad (k = 1, 2, \ldots, n)$$

n'admet que des solutions nulles. Par conséquent

$$u(x_1, x_2, \ldots, x_n) > 0 \text{ pour } |x_1| + |x_2| + \ldots + |x_n| \neq 0.$$

Le théorème est démontré.

§ 14. Méthode de relaxation

Soit le système d'équations linéaires

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{array}\right\} \qquad (1)$$

Transformons ce système de la façon suivante: transposons les termes constants à gauche et divisons la première équation par $-a_{11}$, la deuxième par $-a_{22}$, etc. On obtient ainsi un système *prêt à subir la relaxation*:

$$\left.\begin{array}{l} -x_1 + b_{12}x_2 + \ldots + b_{1n}x_n + c_1 = 0, \\ b_{21}x_1 - x_2 + \ldots + b_{2n}x_n + c_2 = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ b_{n1}x_1 + b_{n2}x_2 + \ldots - x_n + c_n = 0, \end{array}\right\} \qquad (2)$$

où

$$b_{ij} = -\frac{a_{ij}}{a_{ii}} \qquad (i \neq j) \quad \text{et} \quad c_i = \frac{b_i}{a_{ii}}.$$

Soit $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ l'approximation initiale de la résolution du système (2). En portant ces valeurs dans le système (2), on obtient les *résidus*

$$\left.\begin{array}{l} R_1^{(0)} = c_1 - x_1^{(0)} + \sum\limits_{j=2}^{n} b_{1j} x_j^{(0)}, \\ R_2^{(0)} = c_2 - x_2^{(0)} + \sum\limits_{\substack{j=1 \\ j\neq 2}}^{n}{}' \, b_{2j} x_j^{(0)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ R_n^{(0)} = c_n - x_n^{(0)} + \sum\limits_{j=1}^{n-1} b_{nj} x_j^{(0)}. \end{array}\right\} \qquad (3)$$

Si on donne à l'une des inconnues $x_s^{(0)}$ l'accroissement $\delta x_s^{(0)}$, le résidu correspondant $R_s^{(0)}$ diminue de la valeur $\delta x_s^{(0)}$, alors que tous les autres résidus $R_i^{(0)}$ $(i \neq s)$ augmentent de la valeur $b_{is}\delta x_s^{(0)}$. Ainsi, pour annuler le résidu successif $R_s^{(1)}$, il suffit de donner à la grandeur $x_s^{(0)}$ l'accroissement

$$\delta x_s^{(0)} = R_s^{(0)},$$

pour avoir :

$$R_s^{(1)} = 0$$

et

$$R_i^{(1)} = R_i^{(0)} + b_{is}\delta x_s^{(0)} \qquad \text{avec } i \neq s.$$

La *méthode de relaxation* [3], [4] dans sa forme la plus simple consiste à rendre nulle à chaque étape le résidu maximal en module en modifiant la valeur de la composante d'approximation correspondante. Le processus s'arrête lorsque tous les résidus du dernier système transformé s'annulent avec une précision imposée. Nous n'envisageons pas le problème de convergence de ce processus [4].

Exemple. Résoudre le système

$$\left.\begin{aligned} 10x_1 - 2x_2 - 2x_3 &= 6, \\ -x_1 + 10x_2 - 2x_3 &= 7, \\ -x_1 - x_2 + 10x_3 &= 8, \end{aligned}\right\} \qquad (4)$$

par la méthode de relaxation [3] en effectuant les calculs avec deux décimales.

Solution. Ramenons le système (4) à la forme commode pour la relaxation

$$\left.\begin{aligned} -x_1 + 0{,}2x_2 + 0{,}2x_3 + 0{,}6 &= 0, \\ -x_2 + 0{,}1x_1 + 0{,}2x_3 + 0{,}7 &= 0, \\ -x_3 + 0{,}1x_1 + 0{,}1x_2 + 0{,}8 &= 0. \end{aligned}\right\}$$

En choisissant comme approximations initiales de la solution les valeurs nulles

$$x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0,$$

on trouve les résidus correspondants

$$R_1^{(0)} = 0{,}60; \quad R_2^{(0)} = 0{,}70; \quad R_3^{(0)} = 0{,}80.$$

D'après la théorie générale on pose :

$$\delta x_3^{(0)} = 0{,}80.$$

D'où l'on tire les résidus

$$R_1^{(1)} = R_1^{(0)} + 0{,}2 \cdot 0{,}8 = 0{,}60 + 0{,}16 = 0{,}76;$$
$$R_2^{(1)} = R_2^{(0)} + 0{,}2 \cdot 0{,}8 = 0{,}70 + 0{,}16 = 0{,}86;$$
$$R_3^{(1)} = R_3^{(0)} - 0{,}80 = 0.$$

Ensuite, on pose :

$$\delta x_2^{(1)} = 0{,}86,$$

etc. Les résultats correspondants figurent sur le tableau 23.

Tableau 23

Résolution du système linéaire par la méthode de relaxation

	x_1	R_1	x_2	R_2	x_3	R_3
	0	0,60	0	0,70	0	0,80
		0,16		0,16	0,80	−0,80
		0,76		0,86		0
		0,17	0,86	−0,86		0,09
		0,93		0		0,09
	0,93	−0,93		0,09		0,09
		0		0,09		0,18
		0,04		0,04	0,18	−0,18
		0,04		0,13		0
		0,03	0,13	−0,13		0,01
		0,07		0		0,01
	0,07	−0,07		0,01		0,01
		0		0,01		0,02
		0		0	0,02	−0,02
		0		0,01		0
		0	0,01	−0,01		0
		0		0		0
Σ	1,00		1,00		1,00	

En additionnant tous les accroissements $\delta_i^{(k)}$ ($i = 1, 2, 3$; $k = 0, 1, \ldots$), on obtient la solution

$$x_1 = 0 + 0{,}93 + 0{,}07 = 1{,}00;$$

$$x_2 = 0 + 0{,}86 + 0{,}13 + 0{,}01 = 1{,}00;$$

$$x_3 = 0 + 0{,}80 + 0{,}18 + 0{,}02 = 1{,}00.$$

Pour vérifier, portons les valeurs obtenues dans les équations initiales ; dans le cas considéré la solution du système (4) est exacte.

§ 15. Correction des éléments de la matrice inverse approchée

Soit une matrice régulière A ; il faut trouver la matrice inverse A^{-1}. Supposons que nous avons obtenu une valeur approchée de l'inverse $D_0 \approx A^{-1}$. Pour améliorer la précision, on peut utiliser la méthode des approximations successives sous une forme spéciale. Utilisons comme mesure préalable de l'erreur la différence

$$F_0 = E - AD_0.$$

Si $F_0 = 0$, il est évident que $D_0 = A^{-1}$; donc, si les éléments de la matrice F_0 sont petits en module, les matrices A^{-1} et D_0 sont voisines entre elles. Construisons les approximations successives d'après la formule

$$D_k = D_{k-1} + D_{k-1}F_{k-1} \qquad (k = 1, 2, 3, \ldots); \tag{1}$$

l'erreur correspondante s'écrit

$$F_k = E - AD_k.$$

Evaluons la rapidité de la convergence des approximations successives. On a:

$$F_1 = E - AD_1 = E - A(D_0 + D_0F_0) = E - AD_0(E + F_0) =$$
$$= E - (E - F_0)(E + F_0) = E - (E - F_0^2) = F_0^2.$$

D'une façon analogue

$$F_2 = F_1^2 = F_0^4$$

et en général

$$F_k = F_0^{2^k} \qquad (k = 1, 2, 3, \ldots). \tag{2}$$

Montrons que si

$$\| F_0 \| \leqslant q < 1, \tag{3}$$

où $\| F_0 \|$ est une norme canonique quelconque de la matrice F_0 (chapitre VII, § 7), le processus itératif (1) converge, c'est-à-dire

$$\lim_{k\to\infty} D_k = A^{-1}.$$

En effet, la formule (2) entraîne

$$\| F_k \| \leqslant \| F_0 \|^{2^k} \leqslant q^{2^k}.$$

Donc

$$\lim_{k\to\infty} \| F_k \| = 0$$

et

$$\lim_{k\to\infty} F_k = \lim_{k\to\infty} (E - AD_k) = 0$$

ou

$$E - A \lim_{k\to\infty} D_k = 0,$$

soit

$$\lim_{k\to\infty} D_k = A^{-1}E = A^{-1}.$$

Notre proposition est ainsi démontrée.

En particulier, si les éléments de la matrice $F_0 = [f_{ij}]$ vérifient l'inégalité

$$|f_{ij}| \leqslant \frac{q}{n},$$

où n est l'ordre de la matrice et $0 \leqslant q < 1$, l'utilisation de la m-norme (chapitre VII, § 7) montre que le processus itératif (1) est bien convergent.

Supposons respectée l'inégalité (3) pour évaluer l'erreur

$$R_k = \| A^{-1} - D_k \| \leqslant \| A^{-1} \| \, \| E - AD_k \| = \| A^{-1} \| \, \| F_k \| \leqslant \| A^{-1} \| \, q^{2^k}.$$

Comme

$$AD_0 = E - F_0,$$

il vient

$$A^{-1} = D_0 (E - F_0)^{-1} = D_0 (E + F_0 + F_0^2 + \ldots).$$

D'où

$$\| A^{-1} \| \leqslant \| D_0 \| \{ \| E \| + q + q^2 + \ldots \} = \| D_0 \| \left\{ \| E \| + \frac{q}{1-q} \right\}.$$

Pour une m-norme ou une l-norme on a $\| E \| = 1$, et c'est pourquoi

$$\| A^{-1} \| < \frac{\| D_0 \|}{1-q}.$$

Ainsi

$$\| A^{-1} - D_k \| \leqslant \frac{\| D_0 \|}{1-q} \| F_k \| \tag{4}$$

ou

$$\| A^{-1} - D_k \| \leqslant \frac{\| D_0 \|}{1-q} q^{2^k}, \tag{5}$$

où on entend par norme une m-norme ou une l-norme. La formule (4) entraîne que la convergence du processus (1) pour $q \ll 1$ est très rapide.

En pratique, le processus de l'amélioration de la précision des éléments de la matrice inverse s'arrête dès qu'on vérifie l'inégalité

$$\| D_k - D_{k-1} \| \leqslant \varepsilon,$$

où ε est la précision demandée.

E x e m p l e. Corriger les éléments de la matrice inverse approchée obtenue dans l'exemple du § 7, p. 286.

S o l u t i o n. Pour la matrice

$$A = \begin{bmatrix} 1,8 & -3,8 & 0,7 & -3,7 \\ 0,7 & 2,1 & -2,6 & -2,8 \\ 7,3 & 8,1 & 1,7 & -4,9 \\ 1,9 & -4,3 & -4,9 & -4,7 \end{bmatrix}$$

on obtient par la méthode de Gauss l'inverse approchée

$$D_0 = \begin{bmatrix} -0,21121 & -0,46003 & 0,16284 & 0,26956 \\ -0,03533 & 0,16873 & 0,01573 & -0,08920 \\ 0,23030 & 0,04607 & -0,00944 & -0,19885 \\ -0,29316 & -0,38837 & -0,06128 & 0,18513 \end{bmatrix}$$

telle que

$$AD_0 = E - 10^{-3}\cdot \begin{bmatrix} 0,03 & 0,00 & 0,01 & 0,00 \\ 0,25 & 0,03 & 0,02 & 0,39 \\ 8,08 & 10,17 & 0,18 & -0,09 \\ 0,00 & 0,00 & 0,00 & -0,48 \end{bmatrix}.$$

D'où

$$F_0 = E - AD_0 = 10^{-3}\cdot \begin{bmatrix} 0,03 & 0,00 & 0,01 & 0,00 \\ 0,25 & 0,03 & 0,02 & 0,39 \\ 8,08 & 10,17 & 0,18 & -0,09 \\ 0,00 & 0,00 & 0,00 & -0,48 \end{bmatrix}.$$

Pour améliorer encore la précision des éléments de la matrice D_0, on utilise le processus itératif

$$D_{k+1} = D_k + D_k F_k, \quad F_k = E - AD_k \quad (k = 0, 1, 2, \ldots).$$

Comme

$$q = \| F_0 \|_l = 10^{-3}\cdot(0,03 + 10,17) = 1,02\cdot 10^{-2} \ll 1,$$

le processus itératif converge rapidement.

On a :

$$D_0 F_0 = \begin{bmatrix} -0,21121 & -0,46003 & 0,16284 & 0,26956 \\ -0,03533 & 0,16873 & 0,01573 & -0,08920 \\ 0,23030 & 0,04607 & -0,00944 & -0,19885 \\ -0,29316 & -0,38837 & -0,06128 & 0,18513 \end{bmatrix} \times$$

$$\times 10^{-3} \cdot \begin{bmatrix} 0,03 & 0,00 & 0,01 & 0,00 \\ 0,25 & 0,03 & 0,02 & 0,39 \\ 8,08 & 10,17 & 0,18 & -0,09 \\ 0,00 & 0,00 & 0,00 & -0,48 \end{bmatrix} =$$

$$= 10^{-3} \cdot \begin{bmatrix} 1,19 & 1,64 & 0,02 & -0,32 \\ 0,17 & 0,16 & 0,01 & 0,11 \\ -0,06 & -0,09 & 0,00 & 0,11 \\ 0,39 & 0,61 & 0,00 & -0,24 \end{bmatrix}.$$

D'où

$$D_1 = D_0 + D_0F_0 =$$

$$= \begin{bmatrix} -0,21121 & -0,46003 & 0,16284 & 0,26956 \\ -0,03533 & 0,16873 & 0,01573 & -0,08920 \\ 0,23030 & 0,04607 & -0,00944 & -0,19885 \\ -0,29316 & -0,38837 & 0,06128 & 0,18513 \end{bmatrix} +$$

$$+ 10^{-3} \cdot \begin{bmatrix} 1,19 & 1,64 & 0,02 & -0,32 \\ 0,17 & 0,16 & 0,01 & 0,11 \\ -0,06 & -0,09 & 0,00 & 0,11 \\ 0,39 & 0,61 & 0,00 & -0,24 \end{bmatrix} =$$

$$= \begin{bmatrix} -0,21002 & -0,45839 & 0,16286 & 0,26924 \\ -0,03516 & 0,16889 & 0,01574 & -0,08909 \\ 0,23024 & 0,04598 & -0,00944 & -0,19874 \\ -0,29277 & -0,38776 & 0,06128 & 0,18489 \end{bmatrix}.$$

On peut considérer que

$$A^{-1} \approx D_1,$$

car

$$AD_1 = E - 10^{-5} \cdot \begin{bmatrix} 2 & -2 & 1 & 3 \\ 0 & 2 & -1 & 0 \\ 3 & 4 & -5 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

et

$$F_1 = E - AD_1 = 10^{-5} \cdot \begin{bmatrix} 2 & -2 & 1 & 3 \\ 0 & 2 & -1 & 0 \\ 3 & 4 & -5 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

La formule (4) conduit à l'estimation suivante de l'erreur

$$\| A^{-1} - D_1 \|_l \leqslant \frac{\| D_0 \|_l}{1-q} \| F_1 \|_l.$$

Puisque

$$\| D_0 \|_l = 0{,}46003 + 0{,}16873 + 0{,}04607 + 0{,}38837 < 1{,}07$$

et

$$\| F_1 \|_l = 10^{-5} \cdot (2+2+4) = 8 \cdot 10^{-5},$$

on a finalement :

$$\| A^{-1} - D_1 \|_l \leqslant \frac{1{,}07}{1 - 1{,}02 \cdot 10^{-2}} \cdot 8 \cdot 10^{-5} < 9 \cdot 10^{-5}.$$

R e m a r q u e. Le choix d'une matrice inverse approchée peut se faire de façon différente. On utilise entre autres la méthode de partition des matrices en blocs décrite au chapitre VII, § 12.

En conclusion notons qu'il existe actuellement bien d'autres méthodes de résolution des systèmes linéaires d'équations algébriques (méthode de Purcell, d'escalade [6], de Richardson [7],etc.).

BIBLIOGRAPHIE

1. *V. Faddéeva.* Méthodes numériques de l'algèbre linéaire. Gostekhizdat, 1950, chapitre II.
2. *J. Scarborough.* Numerical Mathematical analysis. John Hopkins, 1950.
3. *M. Salvadori.* Numerical methods in engineering. Prentice-Hall, New York, 1952.
4. *E. Beckenbach.* Modern Mathematics for the Engineer. Mc. Graw-Hill, 1956.
5. *K. Smolitski.* Calcul numérique (résumé d'un cours). Académie militaire technique de l'Air Mojaïski de Léningrad, Léningrad, 1960.
6. *D. Faddéev, V. Faddéeva.* Méthodes numériques de l'algèbre linéaire. Fizmatguiz, 1960, chapitre II.
7. *I. Bérézine, N. Jidkov.* Méthodes de calcul. Fizmatguiz, 1959, chapitre VI.

CHAPITRE IX *

CONVERGENCE DES PROCESSUS ITÉRATIFS DES SYSTÈMES D'ÉQUATIONS LINÉAIRES

§ 1. Conditions suffisantes

Soient le système linéaire réduit

$$x = \alpha x + \beta, \tag{1}$$

la matrice

$$\alpha = [\alpha_{ij}]$$

et le vecteur

$$\beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix}$$

donnés et le vecteur recherché

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}.$$

Théorème. *Le processus itératif d'un système linéaire réduit* (1) *converge vers une solution unique si l'une quelconque des normes canoniques de la matrice* α *est inférieure à l'unité, c'est-à-dire pour le processus itératif*

$$x^{(k)} = \beta + \alpha x^{(k-1)} \qquad (k = 1, 2, \ldots)$$

($x^{(0)}$ *étant arbitraire*) *la condition suffisante de convergence est*

$$\| \alpha \| < 1. \tag{2}$$

Démonstration. En partant d'un vecteur arbitraire $x^{(0)}$, on construit la suite des approximations

$$x^{(1)} = \beta + \alpha x^{(0)},$$
$$x^{(2)} = \beta + \alpha x^{(1)},$$
$$\cdots\cdots\cdots$$
$$x^{(k)} = \beta + \alpha x^{(k-1)}.$$

D'où

$$x^{(k)}=(E+\alpha+\alpha^2+\ldots+\alpha^{k-1})\beta+\alpha^k x^{(0)}. \qquad (3)$$

Comme pour $\|\alpha\|<1$ on a $\|\alpha^k\|\to 0$ quand $k\to\infty$, il vient (cf. chapitre VII, § 10)

$$\lim_{k\to\infty}\alpha^k=0$$

et

$$\lim_{k\to\infty}(E+\alpha+\alpha^2+\ldots+\alpha^{k-1})=\sum_{k=0}^{\infty}\alpha^k=(E-\alpha)^{-1}.$$

Donc, en passant à la limite quand $k\to\infty$ dans l'égalité (3), on a:

$$x=\lim_{k\to\infty}x^{(k)}=(E-\alpha)^{-1}\beta. \qquad (4)$$

La convergence de l'itération est ainsi démontrée. De plus l'égalité (4) entraîne:

$$(E-\alpha)\,x=\beta$$

ou

$$x=\alpha x+\beta,$$

c'est-à-dire le vecteur limite x est une solution du système (1). La matrice du système (1) $E-\alpha$ étant régulière, la solution x est unique.

Corollaire 1. Pour le système (1) le processus itératif converge si:

a) $$\|\alpha\|_m=\max_i\sum_{j=1}^{n}|\alpha_{ij}|<1;$$

ou

b) $$\|\alpha\|_l=\max_j\sum_{i=1}^{n}|\alpha_{ij}|<1;$$

ou encore

c) $$\|\alpha\|_k=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|\alpha_{ij}|^2}<1.$$

En particulier, le processus itératif est nécessairement convergent si les éléments de la matrice α vérifient l'inégalité

$$|\alpha_{ij}|<\frac{1}{n},$$

où n est le nombre d'inconnues du système (1).

En effet, a), b) et c) sont les normes canoniques les plus simples de la matrice α.

Corollaire 2. Pour le système

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad (i = 1, 2, \ldots, n) \tag{5}$$

le processus itératif converge si on a les inégalités

a') $$|a_{ii}| > \sum_{j=1}^{n}{}' |a_{ij}| \quad (i = 1, 2, \ldots, n)$$

ou

b') $$|a_{jj}| > \sum_{i=1}^{n}{}' |a_{ij}| \quad (j = 1, 2, \ldots, n),$$

où l'apostrophe affectée au signe de sommation signifie qu'en sommant on omet les valeurs $i = j$, c'est-à-dire la convergence a lieu si les modules des éléments diagonaux de la matrice $A = [a_{ij}]$ du système (1) dépassent pour chaque ligne la somme des modules des éléments non diagonaux de cette ligne, ou pour chaque colonne dépassent la somme des modules des éléments non diagonaux de cette colonne.

En effet, si l'inégalité a') a lieu, il est évident que l'inégalité correspondante a) du corollaire 1 est vérifiée.

Pour démontrer la deuxième proposition posons dans le système (5):

$$x_i = \frac{z_i}{a_{ii}} \quad (i = 1, 2, \ldots, n),$$

où z_i sont de nouvelles inconnues. On obtient alors le système

$$\sum_{j=1}^{n} \frac{a_{ij}}{a_{jj}} z_j = b_i \quad (i = 1, 2, \ldots, n), \tag{5'}$$

pour lequel l'itération converge ou diverge simultanément avec l'itération du système initial (5). En ramenant par le procédé usuel le système (5') à la forme spéciale (1) et en appliquant la condition b) du corollaire 1, on obtient pour le système (5) la condition suffisante de convergence du processus itératif:

$$\sum_{i=1}^{n}{}' \left| \frac{a_{ij}}{a_{jj}} \right| < 1 \quad (j = 1, 2, \ldots, n)$$

ou

$$|a_{jj}| > \sum_{i=1}^{n}{}' |a_{ij}| \quad (j = 1, 2, \ldots, n).$$

§ 2. Estimation de l'erreur des approximations du processus itératif

Soient $x^{(k-1)}$ et $x^{(k)}$ $(k \geqslant 1)$ deux approximations successives de la solution du système linéaire $x = \alpha x + \beta$. Pour $p \geqslant 1$, on a:

$$\| x^{(k+p)} - x^{(k)} \| \leqslant \| x^{(k+1)} - x^{(k)} \| + \\ + \| x^{(k+2)} - x^{(k+1)} \| + \ldots + \| x^{(k+p)} - x^{(k+p-1)} \|. \quad (1)$$

Comme

$$x^{(m+1)} = \alpha x^{(m)} + \beta$$

et

$$x^{(m)} = \alpha x^{(m-1)} + \beta,$$

on a

$$x^{(m+1)} - x^{(m)} = \alpha (x^{(m)} - x^{(m-1)})$$

et donc

$$\| x^{(m+1)} - x^{(m)} \| \leqslant \| \alpha \| \| x^{(m)} - x^{(m-1)} \| \leqslant \\ \leqslant \| \alpha \|^{m-k} \| x^{(k+1)} - x^{(k)} \| \quad \text{pour} \quad m > k \geqslant 1.$$

Il vient de la formule (1)

$$\| x^{(p+k)} - x^{(k)} \| \leqslant \| x^{(k+1)} - x^{(k)} \| + \\ + \| \alpha \| \| x^{(k+1)} - x^{(k)} \| + \ldots + \| \alpha \|^{p-1} \| x^{(k+1)} - x^{(k)} \| \leqslant \\ \leqslant \frac{1}{1 - \| \alpha \|} \| x^{(k+1)} - x^{(k)} \|.$$

En passant à la limite dans la dernière inégalité quand $p \to \infty$, on obtient finalement:

$$\| x - x^{(k)} \| \leqslant \frac{\| x^{(k+1)} - x^{(k)} \|}{1 - \| \alpha \|} \quad (2)$$

pour $k \geqslant 1$, ou

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|}{1 - \| \alpha \|} \| x^{(k)} - x^{(k-1)} \|.$$

Si

$$\| \alpha \| \leqslant \frac{1}{2},$$

la formule précédente se met sous la forme

$$\| x - x^{(k)} \| \leqslant \| x^{(k)} - x^{(k-1)} \|,$$

c'est-à-dire dans ce cas si

$$\| x^{(k)} - x^{(k-1)} \| < \varepsilon,$$

alors également

$$\| x - x^{(k)} \| < \varepsilon.$$

Dans le cas général, si au cours du calcul il s'avère que

$$\| x^{(k)} - x^{(k-1)} \| \leqslant \frac{1-q}{q} \varepsilon,$$

où $q = \| \alpha \| < 1$, alors

$$\| x - x^{(k)} \| \leqslant \varepsilon$$

et par suite

$$| x_i - x_i^{(k)} | \leqslant \varepsilon \quad (i = 1, 2, \ldots, n).$$

Bien entendu, on suppose que les approximations successives $x^{(j)}$ $(j = 0, 1, \ldots, k)$ se calculent exactement sans aucune erreur d'arrondi.

En utilisant les estimations obtenues ci-dessus pour la norme de différence de deux approximations successives, on a suivant la formule (2)

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|^k}{1 - \| \alpha \|} \| x^{(1)} - x^{(0)} \|.$$

En particulier, si l'on choisit

$$x^{(0)} = \beta,$$

il vient

$$x^{(1)} = \alpha\beta + \beta$$

et

$$\| x^{(1)} - x^{(0)} \| = \| \alpha\beta \| \leqslant \| \alpha \| \, \| \beta \|.$$

Par conséquent

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|^{k+1}}{1 - \| \alpha \|} \| \beta \|. \tag{2'}$$

Exemple. Montrer que pour le système

$$\left.\begin{aligned} 10x_1 - x_2 + 2x_3 - 3x_4 &= 0, \\ x_1 + 10x_2 - x_3 + 2x_4 &= 5, \\ 2x_1 + 3x_2 + 20x_3 - x_4 &= -10, \\ 3x_1 + 2x_2 + x_3 + 20x_4 &= 15 \end{aligned}\right\} \tag{3}$$

le processus itératif converge. Combien d'itérations faut-il réaliser pour trouver la solution du système (3) à 10^{-4} près?

Solution. En réduisant le système (3) à la forme spéciale, on obtient:

$$\left.\begin{aligned} x_1 &= 0{,}1x_2 - 0{,}2x_3 + 0{,}3x_4; \\ x_2 &= -0{,}1x_1 + 0{,}1x_3 - 0{,}2x_4 + 0{,}5; \\ x_3 &= -0{,}1x_1 - 0{,}15x_2 + 0{,}05x_4 - 0{,}5; \\ x_4 &= -0{,}15x_1 - 0{,}1x_2 - 0{,}05x_3 + 0{,}75. \end{aligned}\right\} \tag{3'}$$

On en tire la matrice du système

$$\alpha = \begin{bmatrix} 0 & 0,1 & -0,2 & 0,3 \\ -0,1 & 0 & 0,1 & -0,2 \\ -0,1 & -0,15 & 0 & 0,05 \\ -0,15 & -0,1 & -0,05 & 0 \end{bmatrix}.$$

En utilisant, par exemple, la norme $\| \alpha \|_l$, on obtient:

$$\| \alpha \|_l = \max (0,35;\ 0,35;\ 0,35;\ 0,55) = 0,55 < 1.$$

Donc, pour le système (3′) le processus itératif converge.

Prenons pour approximation initiale de la solution x:

$$x^{(0)} = \beta = \begin{bmatrix} 0 \\ 0,5 \\ -0,5 \\ 0,75 \end{bmatrix}.$$

D'où

$$\| \beta \|_l = 0 + 0,5 + 0,5 + 0,75 = 1,75.$$

Soit k le nombre d'itérations, nécessaire pour obtenir la précision donnée. En appliquant la formule (2′) on aura:

$$\| x - x^{(k)} \| \leqslant \frac{\| \alpha \|_l^{k+1} \| \beta \|_l}{1 - \| \alpha \|_l} = \frac{0,55^{k+1} \cdot 1,75}{0,45} < 10^{-4}.$$

Il en résulte

$$0,55^{k+1} < \frac{45}{175} \cdot 10^{-4}$$

et

$$(k+1) \lg 0,55 < \lg 45 - \lg 175 - 4,$$

c'est-à-dire

$$-(k+1) \cdot 0,25964 < 1,65321 - 2,24304 - 4 = -4,58983.$$

Donc

$$k + 1 > \frac{4,58983}{0,25964} \approx 17,7$$

et

$$k > 16,7.$$

On peut poser $k = 17$.

Il est à noter que l'estimation théorique du nombre d'itérations nécessaires pour assurer la précision donnée s'avère en pratique exagérée.

21—01072

§ 3. Première condition suffisante de la convergence du processus de Seidel

Théorème. *Si le système linéaire*

$$x = \alpha x + \beta \tag{1}$$

vérifie la condition

$$\| \alpha \|_m < 1, \tag{2}$$

où

$$\| \alpha \|_m = \max_i \sum_{j=1}^{n} | \alpha_{ij} |,$$

le processus de Seidel pour le système (1) *converge vers une solution unique quel que soit le choix du vecteur initial* $x^{(0)}$.

Démonstration. Soit $x^{(k)} = \{x_1^{(k)}, \ldots, x^{(k)}$ la $k^{\text{ième}}$ approximation du processus de Seidel. On a:

$$x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k-1)} + \beta_i \tag{3}$$

$$(i = 1, 2, \ldots, n; \ k = 1, 2, \ldots).$$

Si la condition (2) est satisfaite, le système (1) admet une solution unique $x = \{x_1, \ldots, x_n\}$ qui peut s'obtenir, par exemple, à l'aide de la méthode itérative simple. On a:

$$x_i = \sum_{j=1}^{n} \alpha_{ij} x_j + \beta_i \quad (i = 1, 2, \ldots). \tag{4}$$

En retranchant l'égalité (3) de l'égalité (4), on obtient:

$$x_i - x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} (x_j - x_j^{(k}) + \sum_{j=i}^{n} \alpha_{ij} (x_j - x_j^{(k-1)});$$

d'où

$$| x_i - x_i^{(k)} | \leqslant \sum_{j=1}^{i-1} | \alpha_{ij} | \, | x_j - x_j^{(k)} | + \sum_{j=i}^{n} | \alpha_{ij} | \, | x_j - x_j^{(k-1)} \tag{5}$$

$$(i = 1, 2, \ldots, n).$$

D'après le sens de la norme adoptée

$$\| x - x^{(k)} \|_m = \max_i | x_i - x_i^{(k)} |,$$

il s'ensuit donc

$$| x_j - x_j^{(k)} | \leqslant \| x - x^{(k)} \|_m$$

$(j = 1, 2, \ldots, n)$. Donc, on déduit de l'inégalité (5):

$$| x_i - x_i^{(k)} | \leqslant p_i \| x - x^{(k)} \|_m + q_i \| x - x^{(k-1)} \|_m, \tag{6}$$

où

$$p_i = \sum_{j=1}^{i-1} |\alpha_{ij}| \quad \text{et} \quad q_i = \sum_{j=i}^{n} |\alpha_{ij}|.$$

Soit $s = s(k)$ la valeur de l'indice i telle que

$$|x_s - x_s^{(k)}| = \max_i |x_i - x_i^{(k)}| = \| x - x^{(k)} \|_m .$$

En posant dans l'inégalité (6) $i = s$, il vient:

$$\| x - x^{(k)} \|_m \leqslant p_s \| x - x^{(k)} \|_m + q_s \| x - x^{(k-1)} \|_m$$

ou

$$\| x - x^{(k)} \|_m \leqslant \frac{q_s}{1 - p_s} \| x - x^{(k-1)} \|_m .$$

D'où

$$\| x - x^{(k)} \|_m \leqslant \mu \| x - x^{(k-1)} \|_m \qquad (7)$$

avec

$$\mu = \max_i \frac{q_i}{1 - p_i} . \qquad (8)$$

Montrons que

$$\mu \leqslant \| \alpha \|_m < 1.$$

En effet, puisque

$$p_i + q_i = \sum_{j=1}^{n} |\alpha_{ij}| \leqslant \| \alpha \|_m < 1,$$

alors

$$q_i \leqslant \| \alpha \|_m - p_i$$

et donc

$$\frac{q_i}{1 - p_i} \leqslant \frac{\| \alpha \|_m - p_i}{1 - p_i} \leqslant \frac{\| \alpha \|_m - p_i \| \alpha \|_m}{1 - p_i} = \| \alpha \|_m .$$

C'est pourquoi

$$\mu = \| \alpha \|_m < 1.$$

L'inégalité (7) entraîne que

$$\| x - x^{(k)} \|_m \leqslant \mu^k \| x - x^{(0)} \|_m ,$$

donc

$$\lim_{k \to \infty} x^{(k)} = x,$$

ce qui démontre la convergence du processus de Seidel vers la solution cherchée.

R e m a r q u e. Comme la méthode itérative simple donne

$$\| x - x^{(k)} \| \leqslant \| \alpha \|_m \| x - x^{(k-1)} \|,$$

tandis que pour la méthode de Seidel on obtient :

$$\| x - x^{(k)} \| \leqslant \mu \| x - x^{(k-1)} \|,$$

où $\mu \leqslant \| \alpha \|_m$, dans les conditions du théorème, la convergence du processus de Seidel est en général quelque peu meilleure que celle du processus itératif simple. La formule (8) amène que dans ce cas, en appliquant la méthode de Seidel, il est commode de disposer le système (1) de façon que la somme des modules des coefficients de la première équation soit la plus petite

$$q_1 = \sum_{j=1}^{n} | \alpha_{1j} |.$$

§ 4. Estimation de l'erreur des approximations du processus de Seidel suivant la m-norme

Soient $x^{(k)}$ et $x^{(k+1)}$ deux itérations successives du processus de Seidel. En appliquant à ces itérations les transformations utilisées pour démontrer le théorème du § 3, on obtient une inégalité analogue à (7) du § 3 :

$$\| x^{(k+1)} - x^{(k)} \|_m \leqslant \mu \| x^{(k)} - x^{(k-1)} \|_m$$

D'où

$$\| x^{(k+p)} - x^{(k)} \|_m \leqslant \| x^{(k+p)} - x^{(k+p-1)} \|_m + \\ + \| x^{(k+p-1)} - x^{(k+p-2)} \|_m + \ldots + \| x^{(k+1)} - x^{(k)} \|_m \leqslant \\ \leqslant \mu^p \| x^{(k)} - x^{(k-1)} \|_m + \mu^{p-1} \| x^{(k)} - x^{(k-1)} \|_m + \ldots \\ \ldots + \mu \| x^{(k)} - x^{(k-1)} \|_m \leqslant \frac{\mu}{1-\mu} \| x^{(k)} - x^{(k-1)} \|_m.$$

Quand $p \to \infty$, on a :

$$\lim_{p \to \infty} x^{(k+p)} = x$$

et donc

$$\| x - x^{(k)} \|_m \leqslant \frac{\mu}{1-\mu} \| x^{(k)} - x^{(k-1)} \|_m,$$

où

$$\mu = \max_i \frac{\sum_{j=i}^{n} | \alpha_{ij} |}{1 - \sum_{j=1}^{i-1} | \alpha_{ij} |} \leqslant \| \alpha \|_m.$$

En particulier, il vient de l'inégalité obtenue:

$$\| x - x^{(k)} \|_m \leqslant \frac{\mu^k}{1-\mu} \| x^{(1)} - x^{(0)} \|_m,$$

c'est-à-dire

$$| x_i - x_i^{(k)} | \leqslant \frac{\mu^k}{1-\mu} \max_j | x_j^{(1)} - x_j^{(0)} | \quad (i = 1, 2, \ldots, n).$$

§ 5. Deuxième condition suffisante de la convergence du processus de Seidel

Théorème. *Si le système linéaire*

$$x = \alpha \cdot x + \beta \tag{1}$$

vérifie la condition

$$\| \alpha \|_l < 1,$$

où

$$\| \alpha \|_l = \max_j \sum_{i=1}^{n} | \alpha_{ij} |$$

le processus de Seidel converge vers une solution unique (1) *quel que soit le choix du vecteur initial.*

Démonstration. Soit

$$x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k-1)} + \beta_i \quad (i = 1, 2, \ldots, n; \; k = 1, 2, \ldots). \tag{2}$$

Pour la solution exacte $x = \{x_1, x_2, \ldots, x_n\}$ qui existe et qui est unique on a:

$$x_i = \sum_{j=1}^{i-1} \alpha_{ij} x_j + \sum_{j=i}^{n} \alpha_{ij} x_j + \beta_i. \tag{3}$$

En retranchant des égalités (3) les égalités correspondantes (2), on obtient:

$$x_i - x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} (x_j - x_j^{(k)}) + \sum_{j=i}^{n} \alpha_{ij} (x_j - x_j^{(k-1)}).$$

D'où

$$| x_i - x_i^{(k)} | \leqslant \sum_{j=1}^{i-1} | \alpha_{ij} | \, | x_j - x_j^{(k)} | + \sum_{j=i}^{n} | \alpha_{ij} | \, | x_j - x_j^{(k-1)} |$$
$$(i = 1, 2, \ldots, n).$$

En sommant les dernières inégalités on aura:

$$\sum_{i=1}^{n} | x_i - x_i^{(k)} | \leqslant \sum_{i=1}^{n} \sum_{j=1}^{i-1} | \alpha_{ij} | \, | x_j - x_j^{(k)} | + \sum_{i=1}^{n} \sum_{j=i}^{n} | \alpha_{ij} | \, | x_j - x_j^{(k-1)} |,$$

ou, en changeant l'ordre de sommation,

$$\sum_{i=1}^{n} |x_i - x_i^{(k)}| \leqslant \sum_{j=1}^{n-1} |x_j - x_j^{(k)}| \sum_{i=j+1}^{n} |\alpha_{ij}| +$$

$$+ \sum_{j=1}^{n} |x_j - x_j^{(k-1)}| \sum_{i=1}^{j} |\alpha_{ij}|. \qquad (4)$$

Posons

$$s_j = \sum_{i=j+1}^{n} |\alpha_{ij}|, \quad t_j = \sum_{i=1}^{j} |\alpha_{ij}| \quad (j = 1, 2, \ldots, n-1)$$

et

$$s_n = 0, \quad t_n = \sum_{i=1}^{n} |\alpha_{ij}|.$$

Il est évident que

$$s_j + t_j = \sum_{i=1}^{n} |\alpha_{ij}| \leqslant \|\alpha\|_l < 1\,;$$

d'où

$$s_j < 1.$$

L'inégalité (4) se met sous la forme

$$\sum_{i=1}^{n} |x_i - x_i^{(k)}| \leqslant \sum_{j=1}^{n} s_j |x_j - x_j^{(k)}| + \sum_{j=1}^{n} t_j |x_j - x_j^{(k-1)}$$

ou

$$\sum_{j=1}^{n} (1 - s_j) |x_j - x_j^{(k)}| \leqslant \sum_{j=1}^{n} t_j |x_j - x_j^{(k-1)}|.$$

Comme

$$t_j \leqslant \|\alpha\|_l - s_j \leqslant \|\alpha\|_l - s_j \|\alpha\|_l = \|\alpha\|_l (1 - s_j), \qquad (5)$$

on a ensuite:

$$\sum_{=1} (1 - s_j) |x_j - x_j^{(k)}| \leqslant \|\alpha\|_l \sum_{i=1}^{n} (1 - s_j) |x_j - x_j^{(k-1)}| \leqslant$$

$$\leqslant \|\alpha\|_l^k \sum_{i=1}^{n} (1 - s_j) |x_j - x_j^{(0)}|. \qquad (6)$$

En passant à la limite lorsque $k \to \infty$ et en tenant compte de ce que $\|\alpha\|_l < 1$, on a:

$$\lim_{k \to \infty} \sum_{j=1}^{n} (1 - s_j) |x_j - x_j^{(k)}| = 0.$$

Par suite

$$\lim_{k \to \infty} x_j^{(k)} = x_j \quad (j = 1, 2, \ldots, n),$$

ce qu'il fallait démontrer.

§ 6. Estimation de l'erreur des approximations du processus de Seidel suivant la l-norme

Soit

$$\sigma_{k+1} = \sum_{j=1}^{n} (1-s_j) \, | x_j^{(k+1)} - x_j^{(k)} | \quad (k=0, 1, 2, \ldots).$$

En appliquant aux deux itérations successives $x_j^{(k)}$ et $x_j^{(k+1)}$ les transformations analogues à celles du paragraphe précédent, on obtient l'inégalité ((6), § 5)

$$\sigma_{k+1} \leqslant \rho \sigma_k, \tag{1}$$

où, en vertu de l'inégalité (5) du § 5,

$$\rho = \max_j \frac{t_j}{1-s_j} \leqslant \| \alpha \|_l.$$

Il en résulte

$$\sigma_{k+p} \leqslant \rho^p \sigma_k \quad (p=1, 2, \ldots).$$

Ensuite, on a:

$$\sum_{j=1}^{n} (1-s_j) \, | x_j^{(k+p)} - x_j^{(k)} | \leqslant \sigma_{k+p} + \sigma_{k+p-1} + \ldots + \sigma_{k+1} \leqslant$$

$$\leqslant \rho^p \sigma_k + \rho^{p-1} \sigma_k + \ldots + \rho \sigma_k \leqslant \frac{\rho \sigma_k}{1-\rho}$$

On en déduit pour $p \to \infty$

$$\sum_{j=1}^{n} (1-s_j) \, | x_j - x_j^{(k)} | \leqslant \frac{\rho \sigma_k}{1-\rho}$$

ou

$$\sum_{j=1}^{n} | x_j - x_j^{(k)} | \leqslant \frac{\rho}{(1-s)(1-\rho)} \sum_{j=1}^{n} | x_j^{(k)} - x_j^{(k-1)} |,$$

avec

$$s = \max_j s_j = \max \sum_{i=j+1}^{n} | \alpha_{ij} |.$$

Puisque la formule (1) entraîne

$$\sigma_k \leqslant \rho^{k-1} \sigma_1$$

l'estimation

$$\| x_j - x_j^{(k)} \|_l = \sum_{j=1}^{n} | x_j - x_j^{(k)} | \leqslant \frac{\rho^k}{(1-s)(1-\rho)} \sigma_1 \leqslant$$

$$\leqslant \frac{\rho^k}{1-s)(1-\rho)} \sum^{n} | x_j^{(1)} - x_j^{(0)} |$$

est également vraie.

§ 7. Troisième condition suffisante de la convergence du processus de Seidel

Théorème. *Si le système linéaire*

$$x = \alpha x + \beta \tag{1}$$

vérifie la condition

$$\| \alpha \|_k < 1,$$

où

$$\| \alpha \|_k + \sqrt{\sum_{i,\,j} | \alpha_{ij} |^2},$$

le processus de Seidel pour le système (1) *converge vers sa solution unique quel que soit le choix du vecteur initial.*

Démonstration. Soit

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{et} \quad x^{(p)} = \begin{bmatrix} x_1^{(p)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(p)} \end{bmatrix}$$

respectivement la solution exacte du système (1) et la p-ième approximation ($p = 0, 1, 2, \ldots$) du processus de Seidel de ce système. On a:

$$x_i = \sum_{j=1}^{i-1} \alpha_{ij} x_j + \sum_{j=i}^{n} \alpha_{ij} x_j + \beta_i$$

et

$$x_i^{(p)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(p)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(p-1)} + \beta_i$$

($i = 1, 2, \ldots, n$). D'où

$$x_i - x_i^{(p)} = \sum_{i=1}^{i-1} \alpha_{ij} (x_j - x_j^{(p)}) + \sum_{j=i}^{n} \alpha_{ij} (x_j - x_j^{(p-1)})$$

et donc

$$| x_i - x_i^{(p)} |^2 \leqslant \Big\{ \sum_{j=1}^{i-1} | \alpha_{ij} | \, | x_j - x_j^{(p)} | + \sum_{j=i}^{n} | \alpha_{ij} | \, | x_j - x_j^{(p-1)} | \Big\}^2.$$

En appliquant l'inégalité de Cauchy (chapitre VII, § 7) à la somme de tous les termes de l'accolade, on a:

$$| x_i - x_i^{(p)} |^2 \leqslant s_i \Big\{ \sum_{j=1}^{i-1} | x_j - x_j^{(p)} |^2 + \sum_{j=i}^{n} | x_j - x_i^{(p-1)} |^2 \Big\} \tag{2}$$

avec

$$s_i = \sum_{j=1}^{n} |\alpha_{ij}|^2 \quad (i = 1, 2, \ldots, n).$$

En effectuant la sommation des inégalités (2) par rapport à i de 1 à n on obtient :

$$\sum_{i=1}^{n} |x_i - x_i^{(p)}|^2 \leqslant \sum_{i=1}^{n} \sum_{j=1}^{i-1} s_i |x_j - x_j^{(p)}|^2 + \sum_{i=1}^{n} \sum_{j=i}^{n} s_i |x_j - x_j^{(p-1)}|^2.$$

Le changement de l'indice de sommation dans le premier membre et de l'ordre de sommation dans le deuxième membre de la dernière inégalité amène

$$\sum_{j=1}^{n} |x_j - x_j^{(p)}|^2 \leqslant \sum_{j=1}^{n-1} |x_j - x_j^{(p)}|^2 \sum_{i=j+1}^{n} s_i + \\ + \sum_{j=1}^{n} |x_j - x_j^{(p-1)}|^2 \sum_{i=1}^{j} s_i. \qquad (3)$$

Soient

$$S_j = \sum_{i=j+1}^{n} s_i, \quad T_j = \sum_{i=1}^{j} s_i \quad (j = 1, 2, \ldots, n-1)$$

et

$$S_n = 0, \quad T_n = \sum_{i=1}^{n} s_i.$$

Il est évident que

$$S_j + T_j = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} \sum_{j=1}^{n} |\alpha_{ij}|^2 = \|\alpha\|_k^2 < 1 \quad (j = 1, 2, \ldots, n). \qquad (4)$$

Utilisant ces notations on peut mettre l'inégalité (3) sous la forme

$$\sum_{j=1}^{n} |x_j - x_j^{(p)}|^2 \leqslant \sum_{j=1}^{n} S_j |x_j - x_j^{(p)}|^2 + \sum_{j=1}^{n} T_j |x_j - x_j^{(p-1)}|^2$$

ou

$$\sum_{j=1}^{n} (1 - S_j) |x_j - x_j^{(p)}|^2 \leqslant \sum_{j=1}^{n} T_j |x_j - x_j^{(p-1)}|^2.$$

En vertu de la formule (4) on obtient :

$$T_j = \|\alpha\|_k^2 - S_j \leqslant \|\alpha\|_k^2 - \|\alpha\|_k^2 S_j = \|\alpha\|_k^2 (1 - S_j)$$

et donc

$$\sum_{j=1}^{n} (1 - S_j) |x_j - x_j^{(p)}|^2 \leqslant \|\alpha\|_k^2 \sum_{j=1}^{n} (1 - S_j) |x_j - x_j^{(p-1)}|^2. \qquad (5)$$

On déduit successivement de l'inégalité (5) pour $p > 1$:

$$\sum_{j=1}^{n} (1-S_j) \, | x_j - x_j^{(p)} |^2 \leqslant (\| \alpha \|_k^2)^p \sum_{j=1}^{n} (1-S_j) \, | x_j - x_j^{(0)} |^2 .$$

Comme $\| \alpha \|_k < 1$, on peut en tirer :

$$\lim_{p \to \infty} \sum_{j=1}^{n} (1-S_j) \, | x_j - x_j^{(p)} |^2 = 0,$$

et, compte tenu du fait que $0 \leqslant S_j < 1$ $(j = 1, 2, \ldots, n)$, on obtient :

$$\lim_{p \to \infty} x_j^{(p)} = x_j \quad (j = 1, 2, \ldots, n),$$

ce qu'il fallait démontrer.

Remarque. L'erreur des itérations $x^{(p)}$ $(p = 1, 2, \ldots)$ est évaluée d'une façon analogue à celle du § 6.

BIBLIOGRAPHIE

1. *V. Faddéeva*. Méthodes numériques de l'algèbre linéaire. Gostekhizdat, Moscou-Léningrad, 1950, chapitre II, §§ 17 et 19.

CHAPITRE X

GÉNÉRALITÉS SUR LA THÉORIE DES ESPACES VECTORIELS

§ 1. Notion de l'espace vectoriel

D é f i n i t i o n. Un ensemble ordonné de n nombres complexes $x = (x_1, x_2, \ldots, x_n)$ s'appelle *point* ou *vecteur* d'un espace de dimension n et les nombres $x_1, x_2, \ldots, x_n$ sont dits *composantes* ou *coordonnées* du vecteur x [1], [2], [3]. Voici quelques exemples de vecteurs:

1) les vecteurs libres dans un plan ou dans un espace tridimensionnel sont respectivement des vecteurs bidimensionnels ou tridimensionnels au sens de la définition ci-dessus;

2) toute solution d'un système quelconque d'équations linéaires à n inconnues est un vecteur de dimension n;

3) si on donne une matrice de m lignes et n colonnes, ses lignes sont des vecteurs de dimension m et ses colonnes des vecteurs de dimension n.

Deux vecteurs $x = (x_1, x_2, \ldots, x_n)$ et $y = (y_1, y_2, \ldots, y_n)$ sont considérés é g a u x si et seulement si leurs coordonnées occupant la même place coïncident, c'est-à-dire si $x_i = y_i$ avec $i = 1, 2, \ldots, n$.

Désignons le vecteur $(0, 0, \ldots, 0)$ par $\mathbf{0}$ et appelons-le *vecteur nul*.

La *somme des vecteurs* $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$ est un vecteur

$$x + y = (x_1 + y_1;\ x_2 + y_2;\ \ldots;\ x_n + y_n)$$

dont les coordonnées sont les sommes des coordonnées correspondantes des vecteurs additionnés. L'addition des vecteurs est c o m m u t a t i v e et a s s o c i a t i v e:

1) $$x + y = y + x;$$

2) $$(x + y) + z = x + (y + z).$$

La différence des vecteurs x et y se définit d'une façon analogue. Le vecteur $-x$ qui satisfait à la condition $(-x) + x = 0$ s'appelle vecteur opposé du vecteur x. On montre aisément que

$$x - y = x + (-y).$$

On appelle *produit du vecteur* $x = (x_1, x_2, \ldots, x_n)$ par le nombre k le vecteur

$$kx = (kx_1, kx_2, \ldots, kx_n).$$

On déduit de cette définition les propriétés suivantes d'un produit de vecteur par un nombre:

1) $$k(x \pm y) = kx \pm ky;$$

2) $$(k \pm l)x = kx \pm lx;$$

3) $$k(lx) = (kl)x;$$

4) $$0x = 0;$$

5) $$1x = x;$$

6) $$(-1)x = -x,$$

où k et l sont des nombres quelconques et x et y, des vecteurs.

Pour les vecteurs x et y la *combinaison linéaire*

$$\alpha x + \beta y$$

(α, β sont des nombres) se définit naturellement tout comme le vecteur de coordonnées $\alpha x_j + \beta y_j$ ($j = 1, 2, \ldots, n$).

Tout ensemble des vecteurs de dimension n, muni des opérations d'addition des vecteurs et de multiplication des vecteurs par un nombre qui ne font pas dépasser les limites de cet ensemble, est dit *espace vectoriel*. En particulier, l'ensemble de tous les vecteurs de dimension n forme un espace vectoriel E_n de dimension n.

§ 2. Dépendance linéaire des vecteurs

Définition 1. Les vecteurs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ de l'espace E_n sont dits *linéairement dépendants* s il existe des nombres $c_1, c_2, \ldots, c_m$ non tous nuls et tels que

$$c_1x^{(1)} + c_2x^{(2)} + \ldots + c_mx^{(m)} = 0. \quad (1)$$

Soit, par exemple, $c_m \neq 0$. L'égalité (1) entraîne

$$x^{(m)} = \gamma_1 x^{(1)} + \gamma_2 x^{(2)} + \ldots + \gamma_{m-1} x^{(m-1)},$$

où

$$\gamma_j = -\frac{c_j}{c_m} \qquad (j = 1, 2, \ldots, m-1).$$

Ainsi, *les vecteurs donnés sont linéairement dépendants si et seulement si l'un d'eux est une combinaison linéaire des autres vecteurs.*

Mais si l'égalité (1) n'est vraie que pour $c_1 = c_2 = \ldots = c_m = 0$, les vecteurs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ sont dits *linéairement indépendants*, *c'est-à-dire que les vecteurs sont linéairement indépendants*

si et seulement si parmi leurs combinaisons linéaires à coefficients non tous nuls il n'y a aucune qui soit un vecteur nul. Notons que parmi les vecteurs linéairement indépendants il ne doit pas y avoir évidemment de vecteur nul.

E x e m p l e 1. Pour le cas d'un espace vectoriel tridimensionnel E_3 la dépendance linéaire de deux vecteurs x et y signifie qu'ils sont parallèles à une certaine droite, et la dépendance linéaire de trois vecteurs x, y et z, qu'ils sont parallèles à un certain plan.

Notons que si une partie des vecteurs est linéairement dépendante, l'ensemble des vecteurs l'est également.

Soit un ensemble des vecteurs

$$x^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}) \quad (j = 1, 2, \ldots, m).$$

Pour déterminer les constantes c_k ($k = 1, 2, \ldots, m$), on obtient en vertu de l'égalité (1) le système

$$\left.\begin{array}{l} c_1 x_1^{(1)} + c_2 x_1^{(2)} + \ldots + c_m x_1^{(m)} = 0, \\ c_1 x_2^{(1)} + c_2 x_2^{(2)} + \ldots + c_m x_2^{(m)} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ c_1 x_n^{(1)} + c_2 x_n^{(2)} + \ldots + c_m x_n^{(m)} = 0. \end{array}\right\} \qquad (2)$$

Si ce système possède des solutions non nulles, les vecteurs donnés sont linéairement dépendants. Dans le cas contraire, ils sont linéairement indépendants.

Considérons la matrice des coordonnées

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ \cdot & \cdot & \cdot & \cdot \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{bmatrix}.$$

Soit r le rang de cette matrice. On montre en algèbre [2] que le système (2) possède des solutions non nulles si et seulement si $r < m$. Les vecteurs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ sont donc linéairement dépendants si $r < m$ et linéairement indépendants si $r = m$ (il est évident que le rang r ne peut pas être supérieur à m).

On en déduit que le rang de la matrice X donne le nombre maximal de vecteurs linéairement indépendants compris dans l'ensemble de vecteurs donné.

De cette façon, si le rang de la matrice X est r, parmi les vecteurs colonnes $x^{(j)}$ ($j = 1, 2, \ldots, m$): 1) il y a r vecteurs linéairement indépendants et 2) tous les $r + 1$ vecteurs ($r + 1 \leqslant m$) de cet ensemble sont linéairement dépendants. Ceci est vrai aussi pour les vecteurs lignes $(x_i^{(1)}, \ldots, x_i^{(m)})$ ($i = 1, 2, \ldots, n$) de la matrice X.

Exemple 2. Etudier la dépendance linéaire du système de vecteurs

$$\begin{aligned} x^{(1)} &= (1, \ -1, \ \ 1, \ -1, \ \ 1); \\ x^{(2)} &= (1, \ \ 0, \ \ 2, \ \ 0, \ \ 1); \\ x^{(3)} &= (1, \ -5, \ -1, \ \ 2, \ -1); \\ x^{(4)} &= (3, \ -6, \ \ 2, \ \ 1, \ \ 1). \end{aligned}$$

Solution. Composons la matrice des coordonnées

$$X = \begin{bmatrix} 1 & 1 & 1 & 3 \\ -1 & 0 & -5 & -6 \\ 1 & 2 & -1 & 2 \\ -1 & 0 & 2 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix}.$$

Pour déterminer le rang r de la matrice X effectuons certaines transformations élémentaires et notamment retranchons de la quatrième colonne la somme des trois premières pour obtenir:

$$X \backsim \begin{bmatrix} 1 & 1 & 1 & 0 \\ -1 & 0 & -5 & 0 \\ 1 & 2 & -1 & 0 \\ -1 & 0 & -2 & 0 \\ 1 & 1 & -1 & 0 \end{bmatrix}.$$

On en déduit que tous les déterminants d'ordre quatre de X sont nuls. Il est clair qu'il y a des mineurs d'ordre trois différents de zéro. Donc $r = 3$, et comme le rang de la matrice est inférieur au nombre de vecteurs, les vecteurs $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$ sont linéairement dépendants. Dans le cas considéré ceci est évident puisque

$$x^{(1)} + x^{(2)} + x^{(3)} - x^{(4)} = 0.$$

Théorème 1. *Le nombre maximal de vecteurs linéairement indépendants d'un espace E_n de dimension n est égal exactement à la dimension de cet espace.*

Démonstration. En premier lieu, l'espace E_n a des systèmes de n vecteurs linéairement indépendants. Tel est, par exemple, l'ensemble de n vecteurs unités:

$$\begin{aligned} e_1 &= (1, \ 0, \ 0, \ \ldots, \ 0); \\ e_2 &= (0, \ 1, \ 0, \ \ldots, \ 0); \\ & \ldots\ldots\ldots\ldots \\ e_n &= (0, \ 0, \ 0, \ \ldots, \ 1). \end{aligned}$$

Si

$$c_1e_1 + c_2e_2 + \ldots + c_ne_n = (c_1, c_2, \ldots, c_n) = 0,$$

il est évident que $c_1 = c_2 = \ldots = c_n = 0$.

Montrons que si le nombre de vecteurs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ est supérieur à n ($m > n$), ils sont nécessairement linéairement dépendants. En effet, la matrice des coordonnées de ces vecteurs est $n \times m$ et, par conséquent, son rang est $r \leqslant \min(n, m) = n < m$. Il en résulte que ces vecteurs sont linéairement dépendants.

Définition 2. Un ensemble quelconque de n vecteurs linéairement indépendants de l'espace de dimension n s'appelle *base* de cet espace.

Théorème 2. *Tout vecteur d'un espace E_n de dimension n peut être représenté d'une seule façon sous forme d'une combinaison linéaire des vecteurs de base.*

Démonstration. Soit $x \in E_n$ et $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ la base de l'espace E_n. En vertu du théorème 1, les vecteurs $x, \varepsilon_1, \varepsilon_2, \ldots$ $\ldots, \varepsilon_n$ sont linéairement dépendants:

$$c_0x + c_1\varepsilon_1 + c_2\varepsilon_2 + \ldots + c_n\varepsilon_n = 0, \tag{3}$$

où un certain coefficient $c_j \neq 0$ $(0 \leqslant j \leqslant n)$.

Dans l'égalité (3), le coefficient $c_0 \neq 0$, car dans le cas contraire on aurait

$$c_1\varepsilon_1 + c_2\varepsilon_2 + \ldots + c_n\varepsilon_n = 0,$$

où $c_j \neq 0$ $(j \geqslant 1)$, ce qui contredit la dépendance linéaire des vecteurs $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Nous pouvons donc résoudre l'égalité (3) par rapport à x:

$$x = \xi_1\varepsilon_1 + \xi_2\varepsilon_2 + \ldots + \xi_n\varepsilon_n, \tag{4}$$

avec

$$\xi_1 = -\frac{c_1}{c_0}, \quad \xi_2 = -\frac{c_2}{c_0}, \quad \ldots, \quad \xi_n = -\frac{c_n}{c_0}.$$

Ainsi, tout vecteur x de l'espace E_n est une combinaison linéaire des vecteurs de la base. Le développement (4) est unique. En effet, s'il existe un autre développement

$$x = \xi_1'\varepsilon_1 + \xi_2'\varepsilon_2 + \ldots + \xi_n'\varepsilon_n \tag{4'}$$

différant du premier, en retranchant l'égalité (4') de l'égalité (4) on obtient:

$$0 = (\xi_1 - \xi_1')\,\varepsilon_1 + (\xi_2 - \xi_2')\,\varepsilon_2 + \ldots + (\xi_n - \xi_n')\,\varepsilon_n, \tag{5}$$

où au moins l'un des coefficients $\xi_j - \xi_j' \neq 0$. L'égalité (5) est impossible du fait que les vecteurs de base sont linéairement indépendants. Par conséquent, il n'existe qu'un seul développement du type (4).

Interprétation géométrique. Pour le cas d'un espace tridimensionnel, la formule (4) est équivalente au développement du vecteur x suivant les directions de trois vecteurs non coplanaires donnés ε_1, ε_2 et ε_3 (fig. 49).

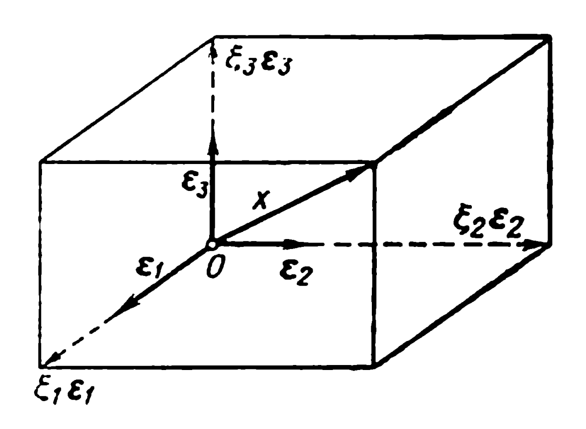


Fig. 49.

D é f i n i t i o n 3. Si ε_1, $\varepsilon_2, \ldots, \varepsilon_n$ est une base d'un espace de dimension n et si

$$x = \xi_1\varepsilon_1 + \xi_2\varepsilon_2 + \ldots + \xi_n\varepsilon_n,$$

les nombres $\xi_1, \xi_2, \ldots, \xi_n$ s'appellent *coordonnées* du vecteur x dans la base donnée $\varepsilon_1, \varepsilon_2, \ldots, \ldots, \varepsilon_n$. Remarquons que les coordonnées du vecteur

$$x = (x_1, x_2, \ldots, x_n)$$

sont ses coordonnées dans la base des vecteurs unités

$$e_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj}) \quad (j = 1, 2, \ldots, n),$$

où δ_{nj} est le symbole de Kronecker. Donc, on a le développement principal

$$x = x_1e_1 + x_2e_2 + \ldots + x_ne_n. \qquad (6)$$

Appelons *base initiale* de l'espace la base des vecteurs unités e_j $(j = 1, 2, \ldots, n)$.

D é f i n i t i o n 4. L'ensemble E_k des vecteurs de l'espace E_n de dimension n s'appelle *sous-espace linéaire* de E_n s'il vérifie les conditions suivantes:

1) $x \in E_k$ et $y \in E_k$ entraînent $x + y \in E_k$;

2) $x \in E_k$ entraîne $\alpha x \in E_k$, où α est un nombre quelconque. En particulier, $0 \in E_k$.

Il s'ensuit que E_k peut être également considéré comme un espace vectoriel. Le nombre maximal de vecteurs linéairement indépendants de E_k est dit *dimension* de ce sous-espace.

Du théorème 1 on déduit que $k \leqslant n$. Ainsi, un espace E_n peut contenir des sous-espaces: E_1 de dimension un, E_2 de dimension deux, etc., jusqu'à E_n de dimension n (l'espace lui-même). Le vecteur nul 0 peut être considéré comme un espace de dimension nulle.

E x e m p l e 3. Dans un espace ordinaire E_3 de dimension trois, le sous-espace E_1 de dimension un est une d r o i t e, le sous-espace E_2 de dimension deux est un p l a n (fig. 50).

T h é o r è m e 3. *Si $z_1, z_2, \ldots, z_k$ sont des vecteurs de l'espace E_n de dimension n, l'ensemble complet des vecteurs*

$$x = a_1z_1 + a_2z_2 + \ldots + a_kz_k, \qquad (7)$$

où a_j *($j = 1, 2, \ldots, k$) sont des nombres arbitraires, est un sous-espace de* E_n, *et si les vecteurs* $z_1, z_2, \ldots, z_k$ *($k \leqslant n$) sont linéairement indépendants, la dimension de ce sous-espace est* k.

Inversement, tout sous-espace E_k *de l'espace* E_n *se confond avec l'ensemble de toutes les combinaisons linéaires des vecteurs linéairement indépendants* $z_1, z_2, \ldots, z_k$ *de ce sous-espace (vecteurs de base).*

D é m o n s t r a t i o n. La première thèse en résulte immédiatement.

Démontrons la deuxième thèse. Soit $x \in E_k$ et x n'est pas une combinaison linéaire des vecteurs de base $z_1, z_2, \ldots, z_k$. Alors il est évident que les vecteurs $x_1, z_1, z_2, \ldots, z_k$ sont linéairement indépendants et de ce fait E_k en possède $k + 1$. Mais ceci est impossible du fait que par hypothèse le nombre maximal de vecteurs linéairement indépendants de E_k est k.

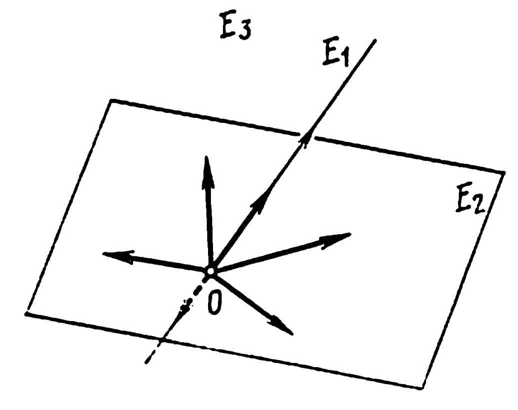


Fig. 50.

Donc un choix arbitraire des nombres $a_1, a_2, \ldots, a_k$ conduit à

$$x = a_1z_1 + a_2z_2 + \ldots + a_kz_k,$$

ce qu'il fallait démontrer.

C o r o l l a i r e. L'ensemble des vecteurs x définis par la formule (7) est le plus petit espace linéaire contenant les vecteurs $z_1, z_2, \ldots, z_k$ (dit *espace engendré par les vecteurs* $z_1, z_2, \ldots, z_k$).

§ 3. Produit scalaire des vecteurs

Soit dans un espace E_n de dimension n les vecteurs

$$x = (x_1, x_2, \ldots, x_n) \quad \text{et} \quad y = (y_1, y_2, \ldots, y_n).$$

Admettons que les coordonnées des vecteurs sont des nombres complexes:

$$x_j = \xi_j + i\xi_j'; \quad y_j = \eta_j + i\eta_j',$$

où $i^2 = -1$; $j = 1, 2, \ldots, n$.

Introduisons des grandeurs conjuguées

$$x_j^* = \xi_j - i\xi_j'; \quad y_j^* = \eta_j - i\eta_j'.$$

Il est évident que

$$x_jx_j^* = |x_j|^2.$$

Par *produit scalaire* de deux vecteurs on entend le nombre

$$(x, y) = \sum_{j=1}^{n} x_j y_j^*. \tag{1}$$

Le produit scalaire jouit des propriétés suivantes:

1. **Définition positive.** Le produit scalaire d'un vecteur par lui-même est un nombre non négatif qui est égal à zéro si et seulement si le vecteur est nul. En effet, la formule (1) donne

$$(x, x) = \sum_{j=1}^{n} x_j x_j^* = \sum_{j=1}^{n} |x_j|^2 \geqslant 0.$$

Il est évident que $(0, 0) = 0$. Inversement, si $(x, x) = 0$, alors $x_j = 0$ $(j = 1, 2, \ldots, n)$ et donc $x = 0$.

2. **Symétrie hermitienne.** Dans la permutation de deux facteurs, le produit scalaire est remplacé par son conjugué. En effet, en appliquant les théorèmes de la grandeur conjuguée d'une somme et de la grandeur conjuguée d'un produit *, on a:

$$(y, x) = \sum_{j=1}^{n} y_j x_j^* = \sum_{j=1}^{n} x_j^* y_j = \left(\sum_{j=1}^{n} x_j y_j^*\right)^* = (x, y)^*.$$

Par suite,

$$(y, x) = (x, y)^*. \tag{2}$$

3. **Le facteur scalaire** qui se trouve en première place peut être sorti de sous le signe du produit scalaire, c'est-à-dire

$$(\alpha x, y) = \alpha (x, y). \tag{3}$$

Cette propriété se déduit immédiatement de la formule (1).

Corollaire. Le facteur scalaire en deuxième place peut être sorti de sous le signe du produit scalaire en le remplaçant par son conjugué. On a:

$$(x, \alpha y) = (\alpha y, x)^* = [\alpha (y, x)]^* = \alpha^* (y, x)^* = \alpha^* (x, y).$$

Ainsi

$$(x, \alpha y) = \alpha^* (x, y).$$

4. **Distributivité.** Si le premier ou le deuxième vecteur constituent une somme de deux vecteurs, le produit scalaire de ce vecteur est égal à la somme des produits scalaires respectifs des termes de ce

---

* Ce sont les théorèmes suivants:

a) la grandeur conjuguée d'une somme est la somme des grandeurs conjuguées de ses termes;

b) la grandeur conjuguée d'un produit est le produit des grandeurs conjugées de ses facteurs.

vecteur. En effet, soit

$$x = x^{(1)} + x^{(2)},$$

où $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})$ $(k = 1, 2)$.

En partant de la définition de la somme des vecteurs, on a d'après la formule (1):

$$(x^{(1)} + x^{(2)}, y) = \sum_{j=1}^{n} (x_j^{(1)} + x_j^{(2)}) y_j^* = \sum_{j=1}^{n} x_j^{(1)} y_j^* + \sum_{j=1}^{n} x_j^{(2)} y_j^* = (x^{(1)}, y) + (x^{(2)}, y),$$

c'est-à-dire

$$(x^{(1)} + x^{(2)}, y) = (x^{(1)}, y) + (x^{(2)}, y). \tag{4}$$

Ensuite

$$(x, y^{(1)} + y^{(2)}) = (y^{(1)} + y^{(2)}, x)^* = (y^{(1)}, x)^* + (y^{(2)}, x)^* = (x, y^{(1)}) + (x, y^{(2)}). \tag{5}$$

Les formules (4) et (5) s'étendent aisément à un nombre fini quelconque de vecteurs, à savoir:

$$\left(\sum_{j=1}^{m} x^{(j)}, \sum_{k=1}^{l} y^{(k)}\right) = \sum_{j=1}^{m} \sum_{k=1}^{l} (x^{(j)}, y^{(k)}).$$

En plus de l'*espace complexe de dimension n*, il est utile de considérer l'*espace réel de dimension n*, c'est-à-dire l'ensemble des vecteurs à coordonnées réelles.

Dans un espace réel de dimension n le produit scalaire est égal à la somme des produits des coordonnées respectives des vecteurs

$$(x, y) = \sum_{j=1}^{n} x_j y_j. \tag{1'}$$

Voici les formulations des propriétés du produit scalaire qui viennent d'être exposées:

1) $(x, x) \geqslant 0$, et si $(x, x) = 0$, alors $x = 0$;
2) $(x, y) = (y, x)$;
3) $(\alpha x, y) = (x, \alpha y) = \alpha (x, y)$ (α est un nombre réel);
4) $(x + y, z) = (x, z) + (y, z)$;
$(x, y + z) = (x, y) + (x, z)$.

Le produit scalaire permet de définir les notions métriques principales dans un espace de dimension n: longueur d'un vecteur et angle entre deux vecteurs.

1. Longueur d'un vecteur. On appelle longueur d'un vecteur dans un espace de dimension n le nombre non négatif

$$|x| = +\sqrt{(x, x)}.$$

Il est clair que cette définition s'accorde avec la notion de longueur d'un vecteur dans un espace à trois dimensions.

2. Angle entre deux vecteurs. On appelle angle φ entre deux vecteurs x et y un angle (0 à 180°) tel que

$$\cos\varphi = \frac{(x, y)}{|x||y|} .$$

Dans un espace à trois dimensions cette définition s'accorde avec l'expression ordinaire de l'angle des vecteurs traduite par un produit scalaire. On peut montrer que l'inégalité

$$|(x, y)| \leqslant |x||y|$$

est vraie [1]. Ainsi, dans un espace réel l'angle des vecteurs est réel.

§ 4. Systèmes orthogonaux des vecteurs

D é f i n i t i o n 1. Deux vecteurs x et y de E_n sont dits *orthogonaux* si leur produit scalaire est nul:

$$(x, y) = 0 \tag{1}$$

Si les vecteurs sont non nuls, l'orthogonalité signifie que leur angle est $\frac{\pi}{2}$. Un vecteur nul est évidemment orthogonal à tout vecteur de l'espace.

Ainsi, l'orthogonalité est une propriété généralisée de la perpendicularité.

D é f i n i t i o n 2. Un système de vecteurs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ s'appelle *orthogonal* si tous ses vecteurs sont orthogonaux deux à deux:

$$(x^{(j)}, x^{(k)}) = 0 \quad \text{avec} \quad j \neq k.$$

Remarquons que si le vecteur $x^{(1)}$ est orthogonal aux vecteurs $x^{(2)}, \ldots, x^{(m)}$, il est également orthogonal à n'importe quelle combinaison linéaire de ces derniers; autrement dit, le vecteur $x^{(1)}$ est orthogonal à l'espace engendré par les vecteurs $x^{(2)}, \ldots, x^{(m)}$. En effet, si

$$(x^{(1)}, x^{(k)}) = 0 \quad \text{pour} \quad k = 2, \ldots, m,$$

on a:

$$\left(x^{(1)}, \sum_{k=2}^{m} c_k x^{(k)}\right) = \sum_{k=2}^{m} c_k^* (x^{(1)}, x^{(k)}) = 0,$$

où $c_2, \ldots, c_m$ sont des constantes arbitraires.

T h é o r è m e. *Les vecteurs non nuls $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ orthogonaux deux à deux sont linéairement indépendants.*

D é m o n s t r a t i o n. En effet, soit

$$c_1x^{(1)}+c_2x^{(2)}+\ldots+c_mx^{(m)}=0. \qquad (2)$$

Le produit scalaire de deux membres de l'égalité (2) par $x^{(1)}$ donne

$$c_1^*(x^{(1)}, x^{(1)})+c_2^*(x^{(1)}, x^{(2)})+\ldots+c_m^*(x^{(1)}, x^{(m)})=0,$$

ou, comme

$(x^{(1)}, x^{(1)})\neq 0$ et $(x^{(1)}, x^{(j)})=0$ pour $j\neq 1$,

on a $c_1^*=0$ et $c_1=0$.

On démontre de même que $c_2 = 0, \ldots, c_m = 0$. Il en résulte que les vecteurs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ sont linéairement indépendants.

C o r o l l a i r e. Dans un espace E_n de dimension n le nombre de vecteurs d'un système orthogonal est égal ou inférieur à n.

D é f i n i t i o n 3. La base $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ de E_n est dite *orthogonale* si les vecteurs de base sont orthogonaux deux à deux:

$$(\varepsilon_j, \varepsilon_k) = 0 \quad \text{si} \quad j\neq k \quad (j, k = 1, 2, \ldots, n).$$

Si de plus les vecteurs ε_j $(j = 1, 2, \ldots, n)$ sont des vecteurs unités, la base orthogonale s'appelle *normale* ou *orthonormale*. Dans ce cas on a:

$$(\varepsilon_j, \varepsilon_k) = \delta_{jk},$$

où δ_{jk} est le symbole de Kronecker.

On voit sans peine qu'une base orthonormale la plus simple d'un espace E_n est le système de vecteurs unités

$$\begin{aligned} e_1 &= (1, 0, 0, \ldots, 0), \\ e_2 &= (0, 1, 0, \ldots, 0), \\ &\cdots\cdots\cdots \\ e_n &= (0, 0, 0, \ldots, 1), \end{aligned}$$

qui forment la base initiale.

Une base orthogonale $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ peut toujours être normée en divisant chacun des vecteurs ε_j par sa longueur. Les nouveaux vecteurs

$$\tilde{\varepsilon}_j^{(0)}=\frac{\varepsilon_j}{\sqrt{(\varepsilon_j, \varepsilon_j)}} \qquad (j=1, 2, \ldots, n)$$

forment une base orthonormale.

Exprimons les coordonnées du vecteur x dans une base orthonormale $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Si

$$x=\xi_1\varepsilon_1+\xi_2\varepsilon_2+\ldots+\xi_n\varepsilon_n, \qquad (3)$$

la postmultiplication scalaire de l'égalité (3) par ε_j amène:

$$\xi_j = (x, \varepsilon_j) \qquad (j = 1, 2, \ldots, n). \tag{4}$$

Par analogie avec l'algèbre vectorielle on peut dire que les *coordonnées d'un vecteur dans une base orthonormale sont égales aux projections du vecteur sur les vecteurs correspondants de la base.*

En élevant au carré l'égalité (3), on a:

$$(x, x) = \left(\sum_{j=1}^{n} \xi_j \varepsilon_j, \sum_{k=1}^{n} \xi_k \varepsilon_k\right) =$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} \xi_j \xi_k^* (\varepsilon_j, \varepsilon_k) = \sum_{k=1}^{n} \xi_j \xi_j^* = \sum_{j=1}^{n} |\xi_j|^2, \tag{5}$$

c'est-à-dire que *le carré de la longueur d'un vecteur est égal à la somme des carrés des modules de ses projections sur les vecteurs de base orthonormaux* (*analogue du théorème de Pythagore*). En particulier, si l'espace E_n est réel, la formule (5) peut s'écrire sans module:

$$(x, x) = \sum_{j=1}^{n} (\xi_j)^2. \tag{5'}$$

§ 5. Transformations des coordonnées d'un vecteur avec changement de base

Soient $e_1, e_2, \ldots, e_n$ et $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ deux bases d'un même espace linéaire E_n. Chaque vecteur de la nouvelle (deuxième) base ε_j est muni dans l'ancienne (première) base e_j de certaines coordonnées $s_{1j}, s_{2j}, \ldots, s_{nj}$ *:

$$\varepsilon_j = s_{1j} e_1 + s_{2j} e_2 + \ldots + s_{nj} e_n \qquad (j = 1, 2, \ldots, n). \tag{1}$$

La matrice régulière $S = [s_{ij}]$ s'appelle *matrice de passage* de l'ancienne base à la nouvelle **. Cette matrice est la transposée de la matrice qui détermine la transformation de la base. Soit x un vecteur donné. Désignons par x_i les coordonnées de ce vecteur dans l'ancienne base et par ξ_i ses coordonnées dans la nouvelle base. Il est évident que

$$x = \sum_{i=1}^{n} x_i e_i = \sum_{j=1}^{n} \xi_j \varepsilon_j.$$

* Pour désigner les coordonnées, on écrit d'abord le numéro de l'ancien vecteur de base suivi du numéro du nouveau vecteur de base.

** Le déterminant $\det S \neq 0$, car dans le cas contraire, les vecteurs $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ seraient linéairement dépendants.

On en tire en portant dans la deuxième somme l'expression (1) pour ε_j:

$$x = \sum_{i=1}^{n} x_i e_i = \sum_{j=1}^{n} \xi_j \sum_{i=1}^{n} s_{ij} e_i = \sum_{i=1}^{n} e_i \sum_{j=1}^{n} s_{ij} \xi_j.$$

Les vecteurs $e_1, e_2, \ldots, e_n$ étant linéairement indépendants, on trouve:

$$x_i = \sum_{j=1}^{n} s_{ij} \xi_j \quad (i = 1, 2, \ldots n). \tag{2}$$

Si l'on désigne

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{et} \quad \xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}^{*},$$

la relation (2) peut s'écrire sous la forme matricielle suivante:

$$x = S\xi, \tag{3}$$

c'est-à-dire que *le vecteur donné par les anciennes coordonnées (dans l'ancienne base) est égal au produit de la matrice de passage S* (ou de la matrice transposée qui donne la nouvelle base) *par le vecteur en nouvelles coordonnées.*

La formule (3) entraîne:

$$\xi = S^{-1}x. \tag{4}$$

Indiquons un cas particulier important analogue à la transformation des coordonnées cartésiennes. Soient l'ancienne base $e_1, e_2, \ldots, e_n$ et la nouvelle base $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ réelles et orthonormales

$$(e_i, e_j) = \delta_{ij} \tag{5}$$

et

$$(\varepsilon_i, \varepsilon_j) = \delta_{ij}, \tag{5'}$$

où δ_{ij} est le symbole de Kronecker.

La formule (1) donne alors

$$s_{ij} = (\varepsilon_j, e_i) \quad (i, j = 1, 2, \ldots, n), \tag{6}$$

c'est-à-dire que les éléments de la matrice de passage S sont des *cosinus directeurs* et peuvent être donnés par le tableau 24.

---

* Autrement dit, nous considérons le vecteur x en nouvelles coordonnées comme un vecteur transformé rapporté à l'ancienne base.

Tableau 24

Cosinus des angles des vecteurs unités de deux bases

Vecteurs unités du nouveau système	Vecteurs unités de l'ancien système			
	e_1	e_2	...	e_n
ε_1	s_{11}	s_{21}	...	s_{n1}
ε_2	s_{12}	s_{22}	...	s_{n2}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
ε_n	s_{1n}	s_{2n}	...	s_{nn}

En portant l'expression (1) dans la formule (5′), on a en vertu des formules (5):

$$(\varepsilon_j, \ \varepsilon_k) = \left(\sum_{i=1}^{n} s_{ij} e_i, \ \sum_{i=1}^{n} s_{ik} e_i\right) = \sum_{i=1}^{n} s_{ij} s_{ik} = \delta_{jk},$$

c'est-à-dire 1) *la somme des produits des cosinus directeurs respectifs de deux axes de coordonnées différents du nouveau système orthonormal s'annule* et 2) *pour tout nouvel axe de coordonnées, la somme des carrés des cosinus directeurs est égale à l'unité.* On en déduit

$$S'S = E, \tag{7}$$

c'est-à-dire que *la matrice de passage d'une base orthonormale à une autre est orthogonale* (pour plus de détails cf. § 6).

§ 6. Matrices orthogonales

D é f i n i t i o n. La matrice réelle A s'appelle *orthogonale* si sa transposée A' est égale à son inverse A^{-1}:

$$A' = A^{-1} \tag{1}$$

ou

$$AA' = A'A = E. \tag{2}$$

Une matrice orthogonale jouit des propriétés suivantes:

1. Ses lignes (colonnes) sont orthogonales deux à deux.

En effet, si $A = [a_{ij}]$, l'égalité (2) entraîne:

$$\sum_{k=1}^{n} a_{ik}a_{jk} = 0 \quad \text{pour} \quad i \neq j$$

et

$$\sum_{k=1}^{n} a_{ki}a_{kj} = 0 \quad \text{pour} \quad i \neq j.$$

2. La somme des carrés des éléments de chaque ligne (colonne) est égale à l'unité.

L'égalité (2) pour $i = j$ donne:

$$\sum_{k=1}^{n} a_{ik}^2 = \sum_{k=1}^{n} a_{ki}^2 = 1.$$

3. Le déterminant est égal à ± 1.

En effet, on a en vertu de l'égalité (2):

$$\det A \det A' = \det E.$$

D'où, comme $\det A' = \det A$ et $\det E = 1$,

$$(\det A)^2 = 1$$

et donc

$$\det A = \pm 1.$$

4. La transposée et l'inverse d'une matrice orthogonale sont aussi des matrices orthogonales. Cette propriété découle directement des formules (1) et (2).

§ 7. Orthogonalisation des matrices

Soit la matrice à éléments réels

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Considérons les colonnes de A comme des vecteurs

$$a^{(j)} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \quad (j = 1, 2, \dots, n).$$

Cette matrice peut donc s'écrire 1072 168 2

$$A = \left[\,\boxed{a^{(1)} | \dots | a^{(n)}}\,\right].$$

Théorème 1. *Toute matrice régulière réelle A peut être mise sous forme de produit d'une matrice aux colonnes orthogonales par une matrice triangulaire supérieure*

$$A = RT,$$

où R est une matrice à colonnes orthogonales et T une matrice triangulaire supérieure à éléments unités diagonaux.

Démonstration. Pour simplifier, démontrons le théorème pour le cas où l'ordre de la matrice est $n = 3$. Toutefois les raisonnements seront d'un caractère général. Soit

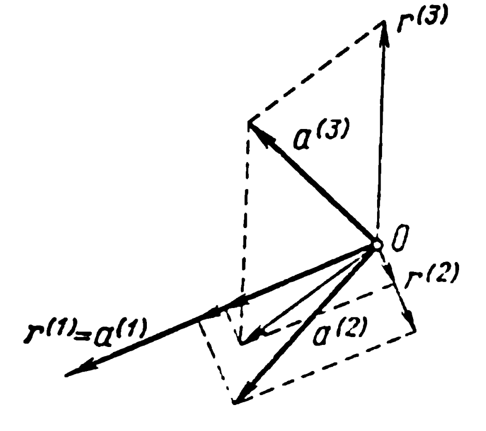


Fig. 51.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Ecrivons cette matrice sous la forme

$$A = [a^{(1)}\ a^{(2)}\ a^{(3)}],$$

où $a^{(j)} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \end{bmatrix}$ sont des vecteurs colonnes.

La matrice A étant régulière, les vecteurs $a^{(1)}$, $a^{(2)}$, $a^{(3)}$ sont linéairement indépendants.

En effet, si ces vecteurs étaient linéairement dépendants, l'une des colonnes de $\det A$ serait une combinaison linéaire de deux autres et par suite $\det A = 0$, ce qui est impossible.

Recherchons la matrice R également sous la forme

$$R = [r^{(1)}\ r^{(2)}\ r^{(3)}],$$

$r^{(j)}$ $(j = 1, 2, 3)$ étant les colonnes orthogonales à obtenir.

Posons

$$r^{(1)} = a^{(1)}. \tag{1}$$

Ensuite, décomposons le vecteur $a^{(2)}$ et $t_{12}r^{(1)}$ et $r^{(2)}$, dont la première composante a la même direction que le vecteur $r^{(1)}$ et la deuxième lui est perpendiculaire (orthogonale) (fig. 51):

$$a^{(2)} = t_{12}r^{(1)} + r^{(2)}, \tag{2}$$

avec

$$(r^{(1)},\ r^{(2)}) = 0. \tag{2'}$$

D'une façon analogue, le vecteur $a^{(3)}$ possède trois composantes $t_{13}r^{(1)}$, $t_{23}r^{(2)}$ et $r^{(3)}$ dont les deux premières sont dirigées respectivement suivant les vecteurs $r^{(1)}$ et $r^{(2)}$, et la dernière est perpendiculaire au vecteur $r^{(1)}$ ainsi qu'au vecteur $r^{(2)}$ (fig. 51):

$$a^{(3)} = t_{13}r^{(1)} + t_{23}r^{(2)} + r^{(3)}, \tag{3}$$

où

$$(r^{(1)}, r^{(3)}) = 0 \quad \text{et} \quad (r^{(2)}, r^{(3)}) = 0. \tag{3'}$$

On voit de la construction que les vecteurs $r^{(1)}$, $r^{(2)}$ et $r^{(3)}$ sont perpendiculaires entre eux. Calculons à partir du système (2) et (3) les vecteurs $r^{(2)}$ et $r^{(3)}$ de même que les coefficients t_{ij}. Le produit scalaire des deux membres de (2) par $r^{(1)} = a^{(1)}$ donne, en vertu de l'orthogonalité (2′),

$$(a^{(2)}, r^{(1)}) = t_{12}(r^{(1)}, r^{(1)});$$

de plus

$$(r^{(1)}, r^{(1)}) \neq 0.$$

Par conséquent,

$$t_{12} = \frac{(a^{(2)}, r^{(1)})}{(r^{(1)}, r^{(1)})}$$

et

$$r^{(2)} = a^{(2)} - t_{12}r^{(1)}.$$

Remarquons que, la matrice A étant régulière, le vecteur $r^{(1)} = a^{(1)} \neq 0$, ce qui amène $(r^{(1)}, r^{(1)}) \neq 0$. Par ailleurs, $r^{(2)} \neq 0$, car autrement les vecteurs $a^{(1)}$ et $a^{(2)}$ seraient linéairement dépendants.

La multiplication scalaire successive analogue des deux membres de l'équation (3) par $r^{(1)}$ et $r^{(2)}$ entraîne, en vertu de l'orthogonalité (2′) et (3′),

$$(a^{(3)}, r^{(1)}) = t_{13}(r^{(1)}, r^{(1)});$$
$$(a^{(3)}, r^{(2)}) = t_{23}(r^{(2)}, r^{(2)}).$$

Il en résulte compte tenu de ce que $(r^{(1)}, r^{(1)}) \neq 0$ et que $(r^{(2)}, r^{(2)}) \neq 0$,

$$t_{13} = \frac{(a^{(3)}, r^{(1)})}{(r^{(1)}, r^{(1)})}, \quad t_{23} = \frac{(a^{(3)}, r^{(2)})}{(r^{(2)}, r^{(2)})}$$

et

$$r^{(3)} = a^{(3)} - t_{13}r^{(1)} - t_{23}r^{(2)}.$$

On vérifie facilement que les vecteurs $r^{(1)}$, $r^{(2)}$ et $r^{(3)}$ construits de cette façon sont orthogonaux deux à deux. Ainsi, on a finalement:

$$\left.\begin{aligned} a^{(1)} &= r^{(1)}, \\ a^{(2)} &= t_{12}r^{(1)} + r^{(2)}, \\ a^{(3)} &= t_{13}r^{(1)} + t_{23}r^{(2)} + r^{(3)}, \end{aligned}\right\} \tag{4}$$

où

$$t_{ij} = \frac{(a^{(j)},\ r^{(i)})}{(r^{(i)},\ r^{(i)})} \quad (i<j)$$

et

$$(r^{(i)},\ r^{(j)}) = 0 \quad \text{pour} \quad i \neq j.$$

Il est évident que le système (4) est équivalent à l'équation matricielle

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \cdot \begin{bmatrix} 1 & t_{12} & t_{13} \\ 0 & 1 & t_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

ou

$$A = RT, \tag{5}$$

$R = [r_{ij}]$ étant la matrice à colonnes orthogonales et $T = [t_{ij}]$, la matrice triangulaire supérieure à diagonale unité.

E x e m p l e. Orthogonaliser les colonnes de la matrice

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 1 \end{bmatrix}.$$

S o l u t i o n. Posons

$$r^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = a^{(1)}.$$

Il vient

$$t_{12} = \frac{(a^{(2)},\ r^{(1)})}{(r^{(1)},\ r^{(1)})} = \frac{1\cdot 0+2\cdot 1+0\cdot 2}{0^2+1^2+2^2} = 0,4.$$

Trouvons maintenant

$$r^{(2)} = a^{(2)} - t_{12}r^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} - 0,4\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1,6 \\ -0,8 \end{bmatrix}.$$

Pour calculer $r^{(3)}$ trouvons t_{13} et t_{23}. On a :

$$t_{13} = \frac{(a^{(3)},\ r^{(1)})}{(r^{(1)},\ r^{(1)})} = \frac{2\cdot 0+0\cdot 1+1\cdot 2}{5} = \frac{2}{5} = 0,4;$$

$$t_{23} = \frac{(a^{(3)},\ r^{(2)})}{(r^{(2)},\ r^{(2)})} = \frac{2\cdot 1+0\cdot 1,6+1\cdot(-0,8)}{1^2+1,6^2+0,8^2} = \frac{1,2}{4,2} \approx 0,3.$$

Il s'ensuit

$$r^{(3)} = a^{(3)} - t_{13}r^{(1)} - t_{23}r^{(2)} = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} - 0,4 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} - 0,3 \begin{bmatrix} 1 \\ 1,6 \\ -0,8 \end{bmatrix} = \begin{bmatrix} 1,70 \\ -0,88 \\ 0,44 \end{bmatrix}.$$

Ainsi,

$$A = \begin{bmatrix} 0 & 1 & 1,7 \\ 1 & 1,6 & -0,88 \\ 2 & -0,8 & 0,44 \end{bmatrix} \begin{bmatrix} 1 & 0,4 & 0,4 \\ 0 & 1 & 0,3 \\ 0 & 0 & 1 \end{bmatrix},$$

les vecteurs

$$r^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}; \quad r^{(2)} = \begin{bmatrix} 1 \\ 1,6 \\ -0,8 \end{bmatrix}; \quad r^{(3)} = \begin{bmatrix} 1,7 \\ -0,88 \\ 0,44 \end{bmatrix}$$

étant orthogonaux deux à deux, ce qu'on peut vérifier par calcul direct.

Dans certains cas il vaut mieux orthogonaliser non pas les colonnes mais les lignes de la matrice en les considérant comme des vecteurs correspondants.

Soit A' la transposée de la matrice donnée A qui est ramenée à la forme

$$A' = RT, \tag{6}$$

avec R matrice aux colonnes orthogonales et T matrice triangulaire supérieure à diagonale unité. En transposant l'égalité (6) on obtient:

$$A = T'R', \tag{7}$$

où T' est une matrice triangulaire inférieure et R' une matrice aux lignes orthogonales. Ainsi le procédé d'orthogonalisation des colonnes d'une matrice décrit ci-dessus convient également pour orthogonaliser les lignes, et nous avons le théorème suivant.

Théorème 2. *Toute matrice régulière réelle peut être représentée sous forme de produit d'une matrice triangulaire inférieure à diagonale unité par une matrice aux lignes orthogonales.*

Indiquons encore un procédé d'orthogonalisation des lignes qui parfois est plus commode en pratique [5]. Soit une matrice régu-

lière réelle

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

Retranchons de chaque i-ème ligne de la matrice A, à partir de la deuxième, sa première ligne multipliée par un certain nombre λ_{i1} $(i = 2, \dots, n)$ assujetti au numéro de la ligne. Il en résulte une matrice transformée

$$A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix},$$

où $a_{ij}^{(1)} = a_{ij}$ pour $i = 1$ et $a_{ij}^{(1)} = a_{ij} - \lambda_{i1} a_{1j}$ pour $i \geqslant 2$.

Choisissons les facteurs λ_{i1} de sorte que la première ligne de la matrice $A^{(1)}$ soit orthogonale à toutes les autres. On a:

$$\sum_{j=1}^{n} a_{1j}^{(1)} a_{ij}^{(1)} = \sum_{j=1}^{n} a_{1j}(a_{ij} - \lambda_{i1} a_{1j}) = \sum_{j=1}^{n} a_{1j} a_{ij} - \lambda_{i1} \sum_{j=1}^{n} a_{1j}^2 = 0.$$

D'où

$$\lambda_{i1} = \frac{\sum_{j=1}^{n} a_{1j} a_{ij}}{\sum_{j=1}^{n} a_{1j}^2} \qquad (i = 2, \dots, n).$$

Soumettons la matrice $A^{(1)}$ à une opération analogue: laissons ses deux premières lignes invariables pour retrancher de toute i-ème ligne, où $i \geqslant 3$, la deuxième ligne de la matrice $A^{(1)}$, multipliée par le nombre λ_{i2} $(i = 3, \dots, n)$. On obtient une nouvelle matrice

$$A^{(2)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & a_{1n}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}^{(2)} & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix},$$

où $a_{ij}^{(2)} = a_{ij}^{(1)}$ pour $i = 1, 2$ et $a_{ij}^{(2)} = a_{ij}^{(1)} - \lambda_{i2} a_{2j}^{(1)}$ pour $i \geqslant 3$.

La première ligne de la matrice $A^{(2)}$ coïncidant avec la première ligne de la matrice $A^{(1)}$ et toutes les autres lignes de la matrice $A^{(2)}$ étant des combinaisons linéaires des lignes de la matrice $A^{(1)}$ orthogonales à la première ligne de $A^{(1)}$, les lignes de la matrice $A^{(2)}$ seront également orthogonales à sa première ligne. Choisissons

les facteurs λ_{i2} de sorte que les lignes de $A^{(2)}$, à partir de la troisième, soient orthogonales à sa deuxième ligne. Il vient:

$$\sum_{j=1}^{n} a_{2j}^{(2)} a_{ij}^{(2)} = \sum_{j=1}^{n} a_{2j}^{(1)} (a_{ij}^{(1)} - \lambda_{i2} a_{2j}^{(1)}) = \sum_{j=1}^{n} a_{2j}^{(1)} a_{ij}^{(1)} - \lambda_{i2} \sum_{j=1}^{n} [a_{2j}^{(1)}]^2 = 0.$$

D'où

$$\lambda_{i2} = \frac{\sum_{j=1}^{n} a_{2j}^{(1)} a_{ij}^{(1)}}{\sum_{j=1}^{n} [a_{2j}^{(1)}]^2} \quad (i = 3, \ldots, n). \tag{A}$$

Ce processus se poursuit jusqu'à ce qu'on obtienne la matrice

$$A^{(n-1)} = \begin{bmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \ldots & a_{1n}^{(n-1)} \\ a_{21}^{(n-1)} & a_{22}^{(n-1)} & \ldots & a_{2n}^{(n-1)} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1}^{(n-1)} & a_{n2}^{(n-1)} & \ldots & a_{nn}^{(n-1)} \end{bmatrix},$$

dont toutes les lignes sont orthogonales deux à deux:

$$\sum_{j=1}^{n} a_{kj}^{(n-1)} a_{ij}^{(n-1)} = 0 \quad \text{pour} \quad k \neq i.$$

La matrice $A^{(n-1)} = \tilde{R}$ aux lignes orthogonales s'obtient à partir de la matrice A après une chaîne d'opérations élémentaires. C'est ce qui justifie l'égalité

$$\tilde{R} = \Lambda A, \tag{8}$$

où Λ est une matrice régulière qui, dans notre cas, est une matrice triangulaire inférieure.

La matrice Λ se rétablit sans peine en soumettant la matrice unité E à toutes les transformations élémentaires subies par la matrice A. La formule (8) donne finalement

$$A = \tilde{T}\tilde{R},$$

$\tilde{T} = \Lambda^{-1}$ étant une matrice triangulaire inférieure.

Indiquons certaines propriétés des matrices aux lignes ou colonnes orthogonales.

L e m m e. *Si les colonnes d'une matrice réelle forment un système de vecteurs orthogonal, le produit de la transposée par la matrice elle-même est égal à la matrice diagonale.*

D é m o n s t r a t i o n. Soit la matrice $A = [a_{ij}]$. Il faut démontrer que $A'A = D$, où $A' = [a_{ji}]$ est la matrice transposée

de A et

$$D = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix}$$

est une matrice diagonale. Posant $D = [d_{ij}]$, on a d'après la règle de multiplication des matrices:

$$d_{ij} = \sum_{k=1}^{n} a_{ki} a_{kj}.$$

Puisque a_{ki} sont les coordonnées du i-ème vecteur $a^{(i)}$ et a_{kj} les coordonnées du j-ième vecteur $a^{(j)}$, on en tire:

$$d_{ij} = \sum_{k=1}^{n} a_{ki} a_{kj} = (a^{(i)}, a^{(j)}) = 0 \text{ si } i \neq j.$$

$D = [d_{ij}]$ est donc une matrice diagonale.

C o r o l l a i r e. Le produit d'une matrice réelle aux lignes orthogonales par sa transposée est égal à une matrice diagonale: $AA' = D$.

T h é o r è m e 3. *Toute matrice réelle régulière A aux colonnes orthogonales est une matrice orthogonale multipliée à droite par une matrice diagonale.*

D é m o n s t r a t i o n. En vertu du lemme, on a:

$$A'A = D, \tag{9}$$

où $D = [d_{ij}]$ est une matrice diagonale. Si $A = [a_{ij}]$, il est évident que

$$d_{ii} = \sum_{k=1}^{n} a_{ki}^2 > 0.$$

Soit

$$\rho_i = \sqrt{d_{ii}} > 0 \qquad (i = 1, 2, \dots, n)$$

et

$$d = \begin{bmatrix} \rho_1 & 0 & \dots & 0 \\ 0 & \rho_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \rho_n \end{bmatrix}.$$

Il est clair que $D = d^2$. La formule (9) entraîne $A'A = d^2$, d'où

$$d^{-1}A'Ad^{-1} = E.$$

Comme

$$(d^{-1})' = d^{-1}, \text{ on a } (Ad^{-1})'(Ad^{-1}) = E.$$

Il en résulte que la matrice $Ad^{-1} = U$ est orthogonale, et donc

$$A = Ud, \tag{10}$$

ce qu'il fallait démontrer.

C o r o l l a i r e. Une matrice réelle régulière aux lignes orthogonales peut être représentée sous forme de produit d'une matrice diagonale par une matrice orthogonale.

En effet, soit A une matrice aux lignes orthogonales; A' est alors une matrice aux colonnes orthogonales. En vertu de la formule (10) on a $A' = Ud$ avec U une matrice orthogonale et d une matrice diagonale qui peut être définie par la relation

$$AA' = d^2.$$

Il en résulte:

$$A = (A')' = d'U' = dU',$$

où U' est également une matrice orthogonale.

R e m a r q u e. Pour rendre orthogonale une matrice réelle régulière A aux colonnes (lignes) orthogonales, il suffit de normer ses colonnes (lignes), c'est-à-dire de diviser tout élément de chaque colonne (ligne) par la racine carrée de la somme des carrés des éléments de cette colonne (ligne). Par exemple, si $A = [a_{ij}]$ est une matrice aux colonnes orthogonales, la matrice

$$\breve{A} = [\breve{a}_{ij}],$$

où

$$\breve{a}_{ij} = \frac{a_{ij}}{\sqrt{\sum_{k=1}^{n} a_{kj}^2}} \qquad (i, j = 1, 2, \ldots, n)$$

est une matrice orthogonale.

§ 8. Application des méthodes d'orthogonalisation à la résolution des systèmes d'équations linéaires

A. P r e m i e r p r o c é d é (orthogonalisation des colonnes) Soit un système linéaire

$$Ax = b \tag{1}$$

à matrice régulière réelle A. En orthogonalisant les colonnes de la matrice A on obtient la matrice R telle que $A = RT$, où T est une matrice triangulaire supérieure. On a:

$$RTx = b. \tag{2}$$

En multipliant à gauche par R' les deux membres de l'égalité (2), on obtient:

$$R'RTx = R'b. \qquad (3)$$

Mais on sait que $R'R = D$, où D est une matrice diagonale. Introduisons la notation $R'b = \beta$ pour obtenir

$$DTx = \beta,$$

d'où

$$x = (DT)^{-1}\beta = T^{-1}D^{-1}\beta. \qquad (4)$$

La matrice D^{-1}, l'inverse de la matrice diagonale, se détermine sans peine; si

$$D = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix},$$

il vient

$$D^{-1} = \begin{bmatrix} d_{11}^{-1} & 0 & \dots & 0 \\ 0 & d_{22}^{-1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn}^{-1} \end{bmatrix}.$$

Il est aussi relativement simple d'obtenir l'inverse T^{-1} de la matrice triangulaire T.

Exemple 1. Résoudre le système

$$\left.\begin{array}{l} 0,4x_1+0,3x_2-0,2x_3=2; \\ 0,6x_1-0,5x_2+0,3x_3=2,5; \\ 0,3x_1+0,2x_2+0,5x_3=11 \end{array}\right\}$$

par orthogonalisation des colonnes.

Solution. Mettons la matrice A du système sous forme de produit d'une matrice R aux colonnes orthogonales par une matrice triangulaire à diagonale unité

$$A = RT = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} 1 & \lambda_{12} & \lambda_{13} \\ 0 & 1 & \lambda_{23} \\ 0 & 0 & 1 \end{bmatrix}.$$

Posons:

$$r^{(1)} = a^{(1)}; \quad r^{(2)} = a^{(2)} - \lambda_{12}r^{(1)}; \quad r^{(3)} = a^{(3)} - \lambda_{13}r^{(1)} - \lambda_{23}r^{(2)}.$$

On a:

$$r^{(1)} = \begin{bmatrix} 0,4 \\ 0,6 \\ 0,3 \end{bmatrix}.$$

D'après les formules (4) du paragraphe précédent on trouve:

$$\lambda_{12}=\frac{(a^{(2)},\ r^{(1)})}{(r^{(1)},\ r^{(1)})}=\frac{0,12-0,3+0,06}{0,16+0,36+0,09}=-\frac{0,12}{0,61}=-0,1967\ ;$$

$$r^{(2)}=\begin{bmatrix}0,3\\-0,5\\0,2\end{bmatrix}+0,1967\begin{bmatrix}0,4\\0,6\\0,3\end{bmatrix}=\begin{bmatrix}0,3787\\-0,3820\\0,2590\end{bmatrix}.$$

Vérification:

$$(r^{(1)},\ r^{(2)})=\begin{bmatrix}0,4\\0,6\\0,3\end{bmatrix}'\begin{bmatrix}0,3787\\-0,3820\\0,2590\end{bmatrix}=\left.\begin{matrix}0,1515\\-0,2292\\0,0777\end{matrix}\right\}=0\ ;$$

$$\lambda_{13}=\frac{(a^{(3)},\ r^{(1)})}{(r^{(1)},\ r^{(1)})}=\frac{-0,08+0,18+0,15}{0,61}=\frac{0,25}{0,61}=0,4098\ ;$$

$$\lambda_{23}=\frac{(a^{(3)},\ r^{(2)})}{(r^{(2)},\ r^{(2)})}=-\frac{0,07574-0,11460+0,12950}{0,35}=-0,1714\ ;$$

$$r^{(3)}=\begin{bmatrix}-0,2\\0,3\\0,5\end{bmatrix}-0,4098\begin{bmatrix}0,4\\0,6\\0,3\end{bmatrix}+0,1714\begin{bmatrix}0,3787\\-0,3820\\0,2590\end{bmatrix}=$$

$$=\begin{bmatrix}-0,2990\\-0,0114\\0,4215\end{bmatrix}.$$

Vérification:

$$(r^{(1)},\ r^{(3)})=(r^{(2)},\ r^{(3)})=0.$$

Ainsi,

$$A=\underbrace{\begin{bmatrix}0,4 & 0,3787 & -0,2990\\0,6 & -0,3820 & -0,0114\\0,3 & 0,2590 & 0,4215\end{bmatrix}}_{R}\underbrace{\begin{bmatrix}1 & -0,1967 & 0,4098\\0 & 1 & -0,1714\\0 & 0 & 1\end{bmatrix}}_{T}.$$

D'après la formule (4), on a:

$$x=T^{-1}D^{-1}R'b,$$

avec $D=R'R$ une matrice diagonale et

$$b=\begin{bmatrix}2\\2,5\\11\end{bmatrix}.$$

Pour la matrice D et son inverse D^{-1} on obtient les valeurs suivantes :

$$D = \begin{bmatrix} 0,61 & 0 & 0 \\ 0 & 0,35 & 0 \\ 0 & 0 & 0,2672 \end{bmatrix} \quad \text{et} \quad D^{-1} = \begin{bmatrix} 1,64 & 0 & 0 \\ 0 & 2,81 & 0 \\ 0 & 0 & 3,75 \end{bmatrix}.$$

Ensuite,

$$R'b = \begin{bmatrix} 0,4 & 0,6 & 0,3 \\ 0,3787 & -0,3820 & 0,2590 \\ -0,2990 & -0,0114 & 0,4215 \end{bmatrix} \begin{bmatrix} 2 \\ 2,5 \\ 11 \end{bmatrix} = \begin{bmatrix} 5,6 \\ 2,67 \\ 4,08 \end{bmatrix}.$$

Enfin, on calcule par le procédé usuel :

$$T^{-1} = \begin{bmatrix} 1 & 0,1967 & -0,3761 \\ 0 & 1 & 0,1714 \\ 0 & 0 & 1 \end{bmatrix}$$

et finalement

$$x = \begin{bmatrix} 1 & 0,1967 & -0,3761 \\ 0 & 1 & 0,1714 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1,64 & 0 & 0 \\ 0 & 2,81 & 0 \\ 0 & 0 & 3,75 \end{bmatrix} \begin{bmatrix} 5,6 \\ 2,67 \\ 4,08 \end{bmatrix} = \begin{bmatrix} 5,0238 \\ 10,0475 \\ 15,0087 \end{bmatrix}.$$

Par conséquent,

$$x_1 = 5,0238; \quad x_2 = 10,0475; \quad x_3 = 15,0087;$$

les valeurs exactes de la solution sont : $x_1 = 5$; $x_2 = 10$; $x_3 = 15$.

B. Deuxième procédé (orthogonalisation des lignes)

Soit le système

$$Ax = b, \tag{5}$$

où $\det A \neq 0$.

Transformons les lignes du système (5) à l'aide du procédé du paragraphe précédent de façon que la matrice A se transforme en matrice R aux lignes orthogonales. Le vecteur b se transformera alors en un certain vecteur β. Il en résulte le système équivalent

$$Rx = \beta. \tag{6}$$

Par suite,

$$x = R^{-1}\beta. \tag{7}$$

On sait que $RR' = D = d^2$, où d est une matrice diagonale et $R = dU$, U étant une matrice orthogonale. Il s'ensuit donc

$$R^{-1} = (dU)^{-1} = U^{-1}d^{-1} = U'd'd^{-2} = (dU)'\,d^{-2} = R'd^{-2} = R'D^{-1}.$$

Ainsi, en vertu de la formule (7) on a finalement :

$$x = R'D^{-1}\beta, \tag{8}$$

avec

$$D = RR'. \tag{9}$$

Utilisant la formule (8) on peut éviter la procédure imposant le plus grand volume de travail pour rechercher l'inverse d'une matrice non diagonale. L'existence de la matrice D^{-1} ne complique pas les calculs du fait que D est une matrice diagonale. La formule (9), nécessaire en fait, peut être utilisée également pour la vérification.

Exemple 2. Résoudre par la méthode d'orthogonalisation des lignes le système

$$\left.\begin{aligned} 3,00x_1 + 0,15x_2 - 0,09x_3 &= 6,00;\\ 0,08x_1 + 4,00x_2 - 0,16x_3 &= 12,00;\\ 0,05x_1 + 0,30x_2 + 5,00x_3 &= 20,00. \end{aligned}\right\} \tag{I}$$

Solution. D'après les formules du paragraphe précédent, déterminons les facteurs:

$$\lambda_{21} = \frac{3,00 \cdot 0,08 + 0,15 \cdot 4,00 + (-0,09) \cdot (-0,16)}{3,00^2 + 0,15^2 + 0,09^2} = \frac{0,8544}{9,0306} = 0,0946;$$

$$\lambda_{31} = \frac{3,00 \cdot 0,05 + 0,15 \cdot 0,30 - 0,09 \cdot 5,00}{3,00^2 + 0,15^2 + 0,09^2} = -\frac{0,2550}{9,0306} = -0,0282.$$

En conservant la première équation du système (I), retranchons de chaque équation suivante la première équation multipliée par les facteurs correspondants λ_{i1} ($i = 2, 3$):

$$\left.\begin{aligned} 3,00x_1 + 0,15x_2 - 0,09x_3 &= 6,00;\\ -0,2038x_1 + 3,9858x_2 - 0,1685x_3 &= 11,4324;\\ 0,1346x_1 + 0,3042x_2 + 4,9975x_3 &= 20,1692. \end{aligned}\right\} \tag{II}$$

Calculons le facteur du système (II)

$$\lambda_{32} = \frac{-0,2038 \cdot 0,1346 + 3,9858 \cdot 0,3042 - 0,1685 \cdot 4,9975}{0,2038^2 + 3,9858^2 + 0,1685^2} = \frac{0,3430}{15,9565} = 0,0215.$$

En conservant les deux premières équations du système (II), retranchons de sa troisième équation la deuxième multipliée par λ_{32}:

$$\left.\begin{aligned} 3,00x_1 + 0,15x_2 - 0,09x_3 &= 6,00;\\ -0,2038x_1 + 3,9858x_2 - 0,1685x_3 &= 11,4324;\\ 0,1390x_1 + 0,2185x_2 + 5,0011x_3 &= 19,9234. \end{aligned}\right\} \tag{III}$$

Les lignes de la matrice

$$R = \begin{bmatrix} 3,00 & 0,15 & -0,09\\ -0,2038 & 3,9858 & -0,1685\\ 0,1390 & 0,2185 & 5,0011 \end{bmatrix}$$

sont orthogonales. Pour vérifier, composons la matrice

$$D = RR' = \begin{bmatrix} 9,0306 & 0,0017 & -0,0002 \\ 0,0017 & 15,9565 & -0,0018 \\ -0,0002 & -0,0018 & 25,0780 \end{bmatrix} \approx$$

$$\approx \begin{bmatrix} 9,0306 & 0 & 0 \\ 0 & 15,9565 & 0 \\ 0 & 0 & 25,0780 \end{bmatrix}.$$

En appliquant la formule (8), on a :

$$x = R'D^{-1}\beta = \begin{bmatrix} 3,00 & -0,2038 & 0,1390 \\ 0,15 & 3,9858 & 0,2185 \\ -0,09 & -0,1685 & 5,0011 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0,1107 & 0 & 0 \\ 0 & 0,0626 & 0 \\ 0 & 0 & 0,0399 \end{bmatrix} \begin{bmatrix} 6,00 \\ 11,4324 \\ 19,9234 \end{bmatrix} = \begin{bmatrix} 1,957 \\ 3,126 \\ 3,803 \end{bmatrix}.$$

Donc

$$x_1 = 1,957; \quad x_2 = 3,126; \quad x_3 = 3,803.$$

C. T r o i s i è m e p r o c é d é (méthode des matrices orthogonales)

Supposons que le système linéaire soit ramené à la forme

$$Rx = \beta, \tag{10}$$

où $R = [r_{ij}]$ est une matrice régulière aux lignes orthogonales et

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$ un vecteur des termes constants.

En multipliant chaque équation de (10) par le normalisateur

$$\mu_i = \frac{1}{\sqrt{\sum_{j=1}^{n} r_{ij}^2}} \qquad (i = 1, 2, \ldots, n),$$

on obtient le système

$$\tilde{R}x = \tilde{\beta}, \tag{11}$$

où $\breve{R} = [\mu_i r_{ij}]$ est une matrice orthogonale et

$$\breve{\beta} = \begin{bmatrix} \mu_1\beta_1 \\ \mu_2\beta_2 \\ \cdots \\ \mu_n\beta_n \end{bmatrix}$$ un nouveau vecteur des termes constants.

L'équation (11) entraîne

$$x = \breve{R}^{-1}\breve{\beta} = \breve{R}'\breve{\beta}. \qquad (12)$$

§ 9. Espace des solutions d'un système homogène

Considérons le système linéaire homogène

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0; \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = 0; \\ \cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = 0 \end{array}\right\} \qquad (1)$$

ou, en abrégé,

$$Ax = 0, \qquad (1')$$

où $A = [a_{ij}]$ et $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ est le vecteur recherché.

Si det $A \neq 0$, en vertu des formules de Cramer le système (1) a la solution unique nulle $x = 0$.

Soit det $A = 0$. Dans ce cas, le système (1) admet une infinité de solutions (y compris les solutions non nulles).

La formule (1') entraîne que 1) si x est une solution de (1'), alors cx, où c est une constante arbitraire, est également une solution de cette équation; 2) si $x^{(1)}$ et $x^{(2)}$ sont des solutions de cette équation, la somme $x^{(1)} + x^{(2)}$ en est également une.

En effet, si

$$Ax = 0,$$

il vient

$$A(cx) = cAx = 0.$$

Par conséquent, cx est une solution du système homogène (1).

D'une façon analogue, si

$$Ax^{(1)} = 0 \quad \text{et} \quad Ax^{(2)} = 0,$$

alors

$$A(x^{(1)} + x^{(2)}) = Ax^{(1)} + Ax^{(2)} = 0 + 0 = 0,$$

c'est-à-dire $x^{(1)} + x^{(2)}$ est également une solution du système homogène (1).

On en tire que toute *combinaison linéaire des solutions d'un système homogène* (1) *est également une solution de ce système.* Il s'ensuit que l'ensemble de toutes les solutions du système homogène (1) forme un espace vectoriel dit *espace des solutions.* Le théorème suivant a lieu.

Théorème. *Si n est le nombre d'inconnues d'un système homogène* (1) *et r le rang de sa matrice A, l'espace des solutions est de dimension* $k = n - r$.

La base de l'espace des solutions s'appelle *famille fondamentale* des solutions. Si l'on connaît pour le système (1) la famille fondamentale des solutions

$$\begin{aligned} x^{(1)} &= (x_1^{(1)},\ x_2^{(1)},\ \ldots,\ x_n^{(1)}),\\ x^{(2)} &= (x_1^{(2)},\ x_2^{(2)},\ \ldots,\ x_n^{(2)}),\\ &\ldots\ldots\ldots\ldots\ldots\\ x^{(k)} &= (x_1^{(k)},\ x_2^{(k)},\ \ldots,\ x_n^{(k)}), \end{aligned}$$

toutes ses solutions sont alors données par la formule

$$x = c_1x^{(1)} + c_2x^{(2)} + \ldots + c_kx^{(k)} \qquad (2)$$

ou, plus en détail

$$\left.\begin{aligned} x_1 &= c_1x_1^{(1)} + c_2x_1^{(2)} + \ldots + c_kx_1^{(k)},\\ x_2 &= c_1x_2^{(1)} + c_2x_2^{(2)} + \ldots + c_kx_2^{(k)},\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ x_n &= c_1x_n^{(1)} + c_2x_n^{(2)} + \ldots + c_kx_n^{(k)}, \end{aligned}\right\}$$

où $c_1, c_2, \ldots, c_k$ sont des constantes arbitraires.

Pour trouver la famille fondamentale des solutions on prélève sur la matrice A le mineur δ_r d'ordre r différent de zéro. Soit

$$\delta_r = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1r}\\ a_{21} & a_{22} & \ldots & a_{2r}\\ \ldots & \ldots & \ldots & \ldots\\ a_{r1} & a_{r2} & \ldots & a_{rr} \end{vmatrix} \neq 0.$$

On peut toujours l'obtenir en permutant les équations de (1) et en changeant la numérotation de ses inconnues. Il est alors aisé de démontrer que les équations du système (1), à partir de la $(r+1)$-ième, sont des conséquences des r premières équations de ce système, c'est-à-dire qu'elles sont vérifiées si les r premières équations du système (1) sont vraies. Il suffit donc de considérer le

sous-système

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1r}x_r = -a_{1,\,r+1}x_{r+1} - \ldots - a_{1n}x_n, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2r}x_r = -a_{2,\,r+1}x_{r+1} - \ldots - a_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rr}x_r = -a_{r,\,r+1}x_{r+1} - \ldots - a_{rn}x_n, \end{array}\right\} \quad (3)$$

dont le déterminant δ_r est différent de zéro.

Dans le système (3), les valeurs des inconnues

$$x_{r+1} = c_1; \quad x_{r+2} = c_2, \ \ldots, \ x_n = c_{n-r} = c_k$$

peuvent être considérées comme arbitraires. En résolvant le système (3) par rapport aux inconnues $x_1, x_2, \ldots, x_r$, on obtient:

$$\left.\begin{array}{l} x_1 = \alpha_{11}c_1 + \alpha_{12}c_2 + \ldots + \alpha_{1k}c_k, \\ x_2 = \alpha_{21}c_1 + \alpha_{22}c_2 + \ldots + \alpha_{2k}c_k, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_r = \alpha_{r1}c_1 + \alpha_{r2}c_2 + \ldots + \alpha_{rk}c_k, \end{array}\right\} \quad (4)$$

où α_{ij} $(i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, k)$ sont des constantes bien définies. D'autre part

$$\left.\begin{array}{l} x_{r+1} = c_1, \\ x_{r+2} = \quad c_2, \\ \ldots\ldots \\ x_n = \quad c_k. \end{array}\right\} \quad (4')$$

Les formules (4) et (4′) donnent le système complet des solutions du système (1). On peut adopter comme famille fondamentale des solutions

$$x^{(1)} = \begin{bmatrix} \alpha_{11} \\ \vdots \\ \alpha_{r1} \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} \alpha_{12} \\ \vdots \\ \alpha_{r2} \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \ldots, \ x^{(k)} = \begin{bmatrix} \alpha_{1k} \\ \vdots \\ \alpha_{rk} \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Ces dernières solutions peuvent s'obtenir directement du système (3) si l'on y pose successivement:

$$\begin{array}{l} x_{r+1} = 1, \quad x_{r+2} = \ldots = x_n = 0; \\ x_{r+1} = 0, \quad x_{r+2} = 1, \quad x_{r+3} = \ldots = x_n = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_{r+1} = \ldots = x_{n-1} = 0, \quad x_n = 1. \end{array}$$

Exemple. Trouver la famille fondamentale des solutions du système homogène

$$\left.\begin{aligned} x_1 - x_2 + 5x_3 - x_4 &= 0, \\ x_1 + x_2 - 2x_3 + 3x_4 &= 0, \\ 3x_1 - x_2 + 8x_3 + x_4 &= 0, \\ x_1 + 3x_2 - 9x_3 + 7x_4 &= 0. \end{aligned}\right\} \qquad (5)$$

Solution. Le rang de la matrice (5) est $r = 2$; de plus

$$\delta = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2 \neq 0.$$

Donc les deux dernières équations du système (5) sont des conséquences des deux premières. On résout le sous-système

$$\left.\begin{aligned} x_1 - x_2 &= -5x_3 + x_4, \\ x_1 + x_2 &= 2x_3 - 3x_4. \end{aligned}\right\}$$

En posant d'abord $x_3 = 1$; $x_4 = 0$, puis $x_3 = 0$ et $x_4 = 1$, on obtient deux solutions linéairement indépendantes

$$x^{(1)} = \left(-\frac{3}{2}, \ \frac{7}{2}, \ 1, \ 0\right);$$

$$x^{(2)} = (-1, \ -2, \ 0, \ 1),$$

qui forment la famille fondamentale des solutions du système (5).

Les vecteurs $x^{(1)}$ et $x^{(2)}$ constituent la base de l'espace des solutions du système donné, et toutes ses solutions sont déterminées par les formules

$$\left.\begin{aligned} x_1 &= -3c_1 - c_2, \\ x_2 &= 7c_1 - 2c_2, \\ x_3 &= 2c_1, \\ x_4 &= c_2, \end{aligned}\right\}$$

où c_1 et c_2 sont des constantes arbitraires (par considérations de commodité, la première constante est mise ici sous la forme $2c_1$).

§ 10. Transformations linéaires

Soient $x_1, x_2, \ldots, x_n$ un système de variables et $y_1, y_2, \ldots, y_n$ un autre système de variables associé au premier par les relations

$$\left.\begin{aligned} y_1 &= f_1(x_1, x_2, \ldots, x_n), \\ y_2 &= f_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots \\ y_n &= f_n(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \qquad (1)$$

où $f_1, f_2, \ldots, f_n$ sont les fonctions données.

Appelons *transformation* le passage du système $x_1, x_2, \ldots, x_n$ au système $y_1, y_2, \ldots, y_n$.

Définition. La transformation (1) est dite *linéaire* si les nouvelles variables $y_1, y_2, \ldots, y_n$ sont des fonctions linéaires homogènes des anciennes variables $x_1, x_2, \ldots, x_n$, c'est-à-dire

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n, \end{aligned}\right\} \qquad (2)$$

les a_{ij} ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n$) étant des constantes.

La transformation linéaire (2) est définie de façon univoque par la matrice carrée de ses coefficients (*matrice de la transformation*)

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}.$$

Inversement, si l'on connaît une matrice carrée $A = [a_{ij}]$, on peut construire la transformation linéaire pour laquelle cette matrice est la matrice des coefficients. Ainsi, on a une correspondance *biunivoque* entre les transformations linéaires et les matrices carrées.

Les systèmes des variables $x_1, x_2, \ldots, x_n$ et $y_1, y_2, \ldots, y_n$ peuvent être considérés comme des vecteurs colonnes

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{et} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

d'un même espace vectoriel E_n. Alors (2) est une application de E_n sur lui-même ou sur sa partie propre.

Exemple 1. La transformation

$$\left.\begin{aligned} y_1 &= x_1 + x_2, \\ y_2 &= x_1 + x_2 \end{aligned}\right\}$$

transforme l'espace E_2 en sous-espace $y_1 = y_2$ de dimension 1.

Les relations (2) sont équivalentes à une relation matricielle

$$y = Ax. \qquad (3)$$

En effet, d'après la règle de multiplication des matrices, on a:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{bmatrix}.$$

On en tire, en vertu de la notion de l'égalité des matrices, les formules (2). D'après la formule (3) la matrice A peut être considérée comme *opérateur* de la transformation linéaire.

Il est facile de voir immédiatement qu'une application linéaire jouit de deux propriétés principales:

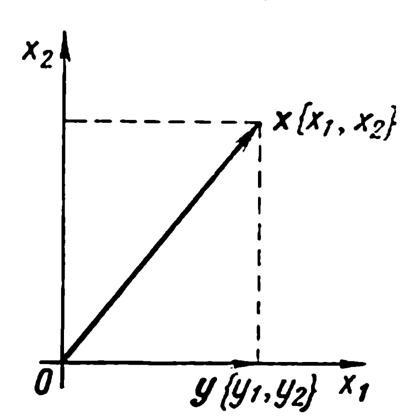


Fig. 52.

1) la constante peut être sortie de sous le signe de l'opérateur:

$$A(\alpha x) = \alpha A x;$$

2) l'opérateur d'une somme de plusieurs vecteurs est une somme des opérateurs de ces vecteurs:

$$A(x + z) = Ax + Az.$$

Il s'ensuit que

$$A(\alpha x + \beta z) = \alpha A x + \beta A z,$$

où x et z sont des vecteurs, α et β des scalaires.

E x e m p l e 2. Supposons que dans le plan Ox_1x_2 on fasse correspondre à tout vecteur

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

le vecteur

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

qui est la projection du vecteur x sur l'axe Ox_1 (*projection*) (fig. 52). Montrer que la transformation donnée est linéaire et trouver sa matrice.

S o l u t i o n. On a évidemment:

$$\left.\begin{aligned} y_1 &= x_1, \\ y_2 &= 0, \end{aligned}\right\}$$

et donc la projection est linéaire. La matrice de la transformation s'écrit

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Elucidons le sens des éléments a_{ij} de la matrice de la transformation A. Considérons les vecteurs unités orientés suivant les axes de coordonnées Ox_1, Ox_2, ..., Ox_n:

$$e_1=\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix},\quad e_2=\begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix},\quad \ldots,\ e_n=\begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}.$$

En appliquant à e_j la transformation A, on aura:

$$Ae_j=\begin{bmatrix}a_{11} & a_{12} & \ldots & a_{1n}\\ a_{21} & a_{22} & \ldots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{n1} & a_{n2} & \ldots & a_{nn}\end{bmatrix}\times\begin{bmatrix}0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix}=\begin{bmatrix}a_{1j}\\a_{2j}\\\vdots\\a_{nj}\end{bmatrix}\qquad (j=1,\ 2,\ \ldots,\ n).$$

Ainsi a_{ij} représente la i-ème coordonnée du transformé du j-ième vecteur unité.

Exemple 3. Supposons que dans le plan Ox_1x_2 tout rayon vecteur x est remplacé par un rayon vecteur y de même longueur, tourné par rapport au premier d'un angle α (*rotation*) (fig. 53).

Montrer que la transformation considérée est linéaire et trouver sa matrice.

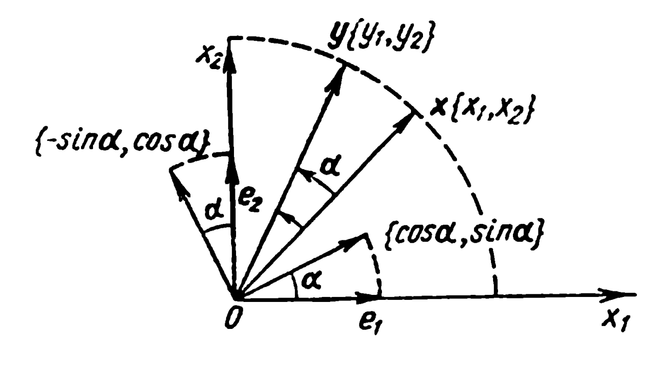


Fig. 53.

Solution. Considérons le deuxième système de référence Oy_1y_2 tourné par rapport au système Ox_1x_2 d'un angle α. Les coordonnées du vecteur y dans le système Oy_1y_2 étant évidemment x_1 et x_2, les coordonnées de ce vecteur dans l'ancien système Ox_1x_2 s'expriment par les formules connues de la géométrie analytique:

$$\left.\begin{aligned} y_1 &= x_1\cos\alpha - x_2\sin\alpha,\\ y_2 &= x_1\sin\alpha + x_2\cos\alpha.\end{aligned}\right\}\qquad (4)$$

Ainsi une rotation est une transformation linéaire et sa matrice s'écrit

$$A=\begin{bmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{bmatrix}.$$

Pour résoudre ce problème on peut procéder d'une autre manière : la rotation fait que le vecteur unité e_1 se transforme évidemment en vecteur $\{\cos\alpha, \sin\alpha\}$, et le vecteur unité e_2 en vecteur $\{-\sin\alpha, \cos\alpha\}$. Il s'ensuit que la matrice de la transformation est

$$A = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix},$$

et nous retrouvons les formules (4).

Définissons les opérations suivantes sur les transformations linéaires A et B: 1) addition $A+B$; 2) multiplication par un scalaire αA; 3) multiplication (composition) BA par conventions naturelles:

1) $$(A+B)\,x = Ax + Bx; \tag{5}$$

2) $$(\alpha A)\,x = \alpha\,(Ax); \tag{5'}$$

3) $$(BA)\,x = B\,(Ax), \tag{5''}$$

où x est un vecteur.

On vérifie aisément que si A et B sont des transformations linéaires, les transformations résultantes $A+B$, αA et BA sont également linéaires.

Le principe suivant a lieu : *à chaque opération* 1)-3) *sur les transformations linéaires est associée une même opération sur leurs matrices*, c'est-à-dire les formules (5), (5'), (5″) peuvent être envisagées comme matricielles.

Pour démontrer cette proposition nous allons nous borner au troisième cas (5″), la vérification des formules (5) et (5') étant plus simple.

Supposons qu'il faut réaliser successivement deux transformations linéaires:

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{aligned}\right\} \tag{6}$$

à matrice $A = [a_{ij}]$ et

$$\left.\begin{aligned} z_1 &= b_{11}y_1 + b_{12}y_2 + \ldots + b_{1n}y_n, \\ z_2 &= b_{21}y_1 + b_{22}y_2 + \ldots + b_{2n}y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ z_n &= b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nn}y_n \end{aligned}\right\} \tag{7}$$

à matrice $B = [b_{ij}]$.

Désignons par $C = [c_{ij}]$ la matrice de la composée de ces transformations dans l'ordre indiqué, c'est-à-dire le passage des variables

$x_1, x_2, \ldots, x_n$ aux variables $z_1, z_2, \ldots, z_n$. En mettant (6) et (7) sous une forme abrégée

$$y_k = \sum_{j=1}^{n} a_{kj}x_j \quad (k=1, 2, \ldots, n), \tag{6'}$$

$$z_i = \sum_{k=1}^{n} b_{ik}y_k \quad (i=1, 2, \ldots, n) \tag{7'}$$

et en portant la formule (6′) dans (7′), on obtient :

$$z_i = \sum_{k=1}^{n} b_{ik}\left(\sum_{j=1}^{n} a_{kj}x_j\right) = \sum_{j=1}^{n} x_j \sum_{k=1}^{n} b_{ik}a_{kj}. \tag{8}$$

Ainsi le coefficient de x_j dans l'expression de z_i, c'est-à-dire l'élément c_{ij} de la matrice C s'écrit

$$c_{ij} = \sum_{k=1}^{n} b_{ik}a_{kj} = b_{i1}a_{1j} + b_{i2}a_{2j} + \ldots + b_{in}a_{nj}.$$

On voit que l'élément de la matrice C qui figure dans la i-ème ligne et j-ième colonne est égal à la somme des produits des éléments correspondants de la i-ème ligne de la matrice B et de la j-ième colonne de la matrice A, c'est-à-dire coïncide avec l'élément respectif du produit de B par A. Par conséquent, $C = BA$.

Si l'on utilise une écriture matricielle, la démonstration devient nettement plus simple. Soient

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{et} \quad z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

les vecteurs correspondants. Les formules (6) et (7) donnent

$$y = Ax \quad \text{et} \quad z = By.$$

D'où

$$z = B(Ax) = (BA)x.$$

La matrice de la transformation résultante est donc $C = BA$.

E x e m p l e 4. Trouver le résultat de la réalisation successive des transformations linéaires

$$y_1 = 5x_1 - x_2 + 3x_3;$$
$$y_2 = x_1 - 2x_2;$$
$$y_3 = 7x_2 - x_3$$

et

$$z_1 = 2y_1 + y_3;$$
$$z_2 = y_2 - 5y_3;$$
$$z_3 = 2y_2.$$

S o l u t i o n. Ecrivons les matrices des transformations

$$A = \begin{bmatrix} 5 & -1 & 3 \\ 1 & -2 & 0 \\ 0 & 7 & -1 \end{bmatrix} \quad \text{et} \quad B = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & -5 \\ 0 & 2 & 0 \end{bmatrix}.$$

Le produit de ces matrices donne la matrice de la transformation résultante

$$C = BA = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & -5 \\ 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} 5 & -1 & 3 \\ 1 & -2 & 0 \\ 0 & 7 & -1 \end{bmatrix} = \begin{bmatrix} 10 & 5 & 5 \\ 1 & -37 & 5 \\ 2 & -4 & 0 \end{bmatrix}.$$

Par conséquent, la transformation linéaire recherchée s'écrit

$$z_1 = 10x_1 + 5x_2 - 5x_3;$$
$$z_2 = x_1 - 37x_2 + 5x_3;$$
$$z_3 = 2x_1 - 4x_2.$$

§ 11. Transformation inverse

Soit la transformation linéaire

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{aligned}\right\} \qquad (1)$$

qui associe le système de variables $x_1, x_2, \ldots, x_n$ au système de variables $y_1, y_2, \ldots, y_n$.

D é f i n i t i o n. Une transformation qui fait correspondre au système de variables $y_1, y_2, \ldots, y_n$ le système de variables $x_1, x_2, \ldots, x_n$ est dite *inverse* par rapport à (1).

Les formules de la transformation inverse s'obtiennent en résolvant le système (1) par rapport aux variables $x_1, x_2, \ldots, x_n$.

Soient A_{ij} $(i, j = 1, 2, \ldots, n)$ les cofacteurs des éléments a_{ij} de la matrice $A = [a_{ij}]$ et

$$\det A = \det [a_{ij}] \doteq \Delta \neq 0.$$

En multipliant les équations de (1) respectivement par A_{11}, A_{21}, . . ., A_{n1} et en additionnant, on a en vertu de la formule connue

$$A_{11}y_1 + A_{21}y_2 + \ldots + A_{n1}y_n = \Delta x_1.$$

D'une façon analogue on déduit:

$$A_{12}y_1 + A_{22}y_2 + \ldots + A_{n2}y_n = \Delta x_2;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$A_{1n}y_1 + A_{2n}y_2 + \ldots + A_{nn}y_n = \Delta x_n.$$

D'où

$$\left.\begin{aligned} x_1 &= \frac{A_{11}}{\Delta} y_1 + \frac{A_{21}}{\Delta} y_2 + \ldots + \frac{A_{n1}}{\Delta} y_n, \\ x_2 &= \frac{A_{12}}{\Delta} y_1 + \frac{A_{22}}{\Delta} y_2 + \ldots + \frac{A_{n2}}{\Delta} y_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= \frac{A_{1n}}{\Delta} y_1 + \frac{A_{2n}}{\Delta} y_2 + \ldots + \frac{A_{nn}}{\Delta} y_n. \end{aligned}\right\} \quad (2)$$

Ainsi l'inverse d'une transformation linéaire est également linéaire (si elle existe).

Théorème. *Une transformation linéaire possède une transformation inverse univoque si et seulement si la matrice de la transformation donnée est régulière. L'inverse d'une transformation linéaire est linéaire et sa matrice est l'inverse de la matrice de la transformation initiale.*

Démonstration. Si $A = [a_{ij}]$ est la matrice de la transformation (1) et $\Delta = \det A \neq 0$, la transformation inverse existe et est définie par les formules (2). La matrice de la transformation inverse s'écrit évidemment

$$\left(\frac{A_{ji}}{\Delta}\right) = A^{-1}.$$

Si $\Delta = 0$, on sait de l'algèbre que le système (1) est soit incompatible, soit indéterminé par rapport aux variables $x_1, x_2, \ldots, x_n$. Il n'existe donc pas de transformation inverse univoque et, de plus, il y a nécessairement des valeurs des variables $y_1, y_2, \ldots, y_n$ pour lesquelles il n'existe pas de valeurs correspondantes des variables $x_1, x_2, \ldots, x_n$. Dans ce cas la transformation est dite *dégénérée*.

Remarque 1. Mettons la transformation (1) sous une forme matricielle

$$y = Ax, \quad (3)$$

$A = [a_{ij}]$ étant la matrice de la transformation; x et y les vecteurs colonnes.

24—01072

Si la transformation A est *non dégénérée* ($\det A \neq 0$), il existe la transformation inverse

$$x = A^{-1}y, \tag{4}$$

et, en vertu de la formule (3) à tout vecteur x de l'espace $Ox_1x_2 \ldots$ $\ldots x_n$ de dimension n correspond un et un seul vecteur y de cet espace, c'est-à-dire que la formule (3) applique l'espace $Ox_1x_2 \ldots x_n$ sur lui-même.

Si la transformation A est *dégénérée* ($\det A = 0$), la formule (3) transforme l'espace $Ox_1x_2 \ldots x_n$ en sous-espace d'un plus petit nombre de dimensions.

E x e m p l e. Considérons la projection (§ 10, exemple 2) définie par la matrice

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Ici A est singulière et la transformation $y = Ax$ associe l'espace Ox_1x_2 à l'axe de coordonnées Ox_1.

R e m a r q u e 2. Convenons d'entendre par Ex une transformation identique qui laisse invariable le vecteur x.

Puisque les relations

$$y = Ax \text{ et } x = A^{-1}y$$

entraînent

$$y = AA^{-1}y \text{ et } x = A^{-1}Ax,$$

il vient

$$AA^{-1} = A^{-1}A = E.$$

§ 12. Vecteurs propres et valeurs propres d'une matrice

Soit la matrice carrée $A = [a_{ij}]$. Considérons la transformation linéaire

$$y = Ax, \tag{1}$$

où x et y sont des vecteurs de dimension n (matrices colonnes) d'un certain espace de dimension n, en général complexe.

D é f i n i t i o n 1. Un vecteur non nul s'appelle *vecteur propre* de la matrice donnée (ou de la transformation linéaire qu'elle définit) si son image par l'application linéaire correspondante est colinéaire à ce vecteur, c'est-à-dire si le vecteur transformé ne se distingue du vecteur initial que par un scalaire.

Autrement dit, le vecteur $x \neq 0$ s'appelle vecteur *propre* de la matrice A si cette matrice transforme le vecteur x en vecteur

$$Ax = \lambda x. \tag{2}$$

Le nombre λ qui figure dans l'égalité (2) est dit *valeur propre* ou *nombre caractéristique* de la matrice A, qui correspond au vecteur propre x donné.

E x e m p l e 1. Considérons la projection dans l'espace bidimensionnel Ox_1x_2, déterminée par la matrice

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Ici les vecteurs propres sont 1) les vecteurs non nuls x dirigés suivant l'axe Ox_1 à valeur propre $\lambda_1 = 1$ et 2) les vecteurs non nuls y dirigés suivant l'axe Ox_2 à valeur propre $\lambda_2 = 0$ (fig. 54).

T h é o r è m e 1. *Dans un espace vectoriel complexe toute transformation linéaire (matrice) possède au moins un vecteur propre réel ou complexe.*

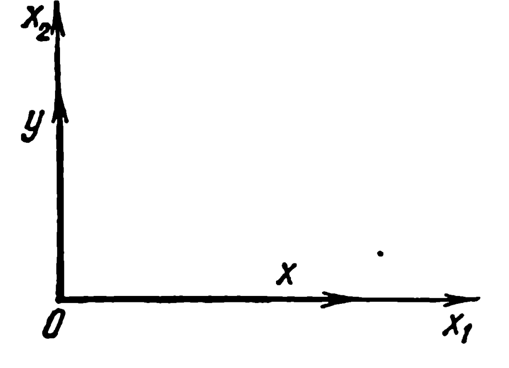


Fig. 54.

D é m o n s t r a t i o n. Soit A une matrice de la transformation linéaire. Les vecteurs propres de A sont des solutions non nulles de l'équation matricielle

$$Ax = \lambda x$$

ou

$$(A - \lambda E)\, x = 0 \qquad (3)$$

avec la matrice $A - \lambda E$, dite *matrice caractéristique*. L'équation (3) est un système linéaire homogène qui a des solutions non nulles si et seulement si le déterminant du système est nul, c'est-à-dire si la condition

$$\det (A - \lambda E) = 0 \qquad (4)$$

est vraie.

Le déterminant (4) est appelé déterminant *caractéristique* (*séculaire*) de la matrice A, et l'équation (4) est dite équation *caractéristique* (*séculaire*) de la matrice A. Sous une forme développée, l'équation caractéristique (4) s'écrit:

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda \end{vmatrix} = 0 \qquad (4')$$

ou

$$\lambda^n - \sigma_1\lambda^{n-1} + \sigma_2\lambda^{n-2} - \ldots + (-1)^{n-1}\sigma_{n-1}\lambda + (-1)^n\sigma_n = 0. \qquad (5)$$

Le polynôme du premier membre de l'équation (5) est un *polynôme caractéristique* de la matrice A. Ses coefficients σ_i ($i = 1, 2, \ldots, n$) se déterminent d'après les règles suivantes. Le coefficient σ_1 est la somme des éléments diagonaux de la matrice A, c'est-à-dire $\sigma_1 = \sum_{i=1}^{n} a_{ii}$. Ce nombre s'appelle *trace* de la matrice A et se note $\sigma_1 = \mathrm{Sp}\, A$. Le coefficient σ_2 est la somme de tous les mineurs diagonaux d'ordre 2 de la matrice A. Dans le cas général, le coefficient σ_k est la somme de tous les mineurs diagonaux d'ordre k de la matrice A. Enfin le terme constant σ_n est égal au déterminant de la matrice A :

$$\sigma_n = \det A.$$

L'équation caractéristique (5) est une équation algébrique de degré n par rapport à λ et possède, comme on le montre en algèbre, au moins une racine réelle ou complexe. Soient $\lambda_1, \lambda_2, \ldots, \lambda_m$ ($m \leqslant n$) les racines distinctes de (5). Ces racines sont les *valeurs propres* ou les *nombres caractéristiques* de la matrice A et l'ensemble de toutes les valeurs propres s'appelle *spectre* de A. Prenons une racine quelconque $\lambda = \lambda_j$ et portons-la dans l'équation (4). On a alors $(A - \lambda_j E)\, x = 0$, ou sous une forme développée

$$\left.\begin{aligned} &(a_{11}-\lambda_j)\, x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0, \\ &a_{21}x_1 + (a_{22}-\lambda_j)\, x_2 + \ldots + a_{2n}x_n = 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn}-\lambda_j)\, x_n = 0. \end{aligned}\right\} \tag{6}$$

Le déterminant du système (6) étant $\det (A - \lambda_j E) = 0$, ce système a forcément des solutions non nulles qui sont précisément les valeurs propres de la matrice A associées à la valeur propre λ_j. Si le rang de la matrice $A - \lambda_j E$ est r ($r < n$), il existe $k = n - r$ vecteurs propres linéairement indépendants

$$x^{(1j)},\ x^{(2j)},\ \ldots,\ x^{(kj)},$$

qui correspondent à la racine λ_j. Le théorème est ainsi démontré.

Remarque. On peut montrer que le nombre de vecteurs propres linéairement indépendants associés à la même racine d'une équation caractéristique ne dépasse pas l'ordre de multiplicité de cette racine. Il s'ensuit, en particulier, que si les racines de l'équation caractéristique (5) sont distinctes, à chaque valeur propre correspond, à un coefficient de proportionnalité près, un vecteur propre et un seul.

Exemple 2. Trouver les valeurs propres et les vecteurs propres de la matrice

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Solution. Composons l'équation caractéristique de A:

$$\begin{bmatrix} 2-\lambda & 1 & 1 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & 2-\lambda \end{bmatrix} = 0.$$

D'où $(\lambda - 1)^2 (4 - \lambda) = 0$ et $\lambda_1 = \lambda_2 = 1$; $\lambda_3 = 4$.

Prenons $\lambda_1 = 1$ et portons-la dans l'équation

$$(A - \lambda_j E)\, x = 0. \tag{7}$$

On a:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

ou

$$\left.\begin{aligned} x_1 + x_2 + x_3 = 0, \\ x_1 + x_2 + x_3 = 0, \\ x_1 + x_2 + x_3 = 0. \end{aligned}\right\} \tag{8}$$

Le rang de la matrice du système (8) étant $r = 1$, deux de ses équations se déduisent de la troisième (ce qui d'ailleurs est évident). Il suffit donc de résoudre l'équation

$$x_1 + x_2 + x_3 = 0.$$

En posant $x_1 = c_1$; $x_2 = c_2$, on obtient:

$$x_3 = -(c_1 + c_2),$$

où c_1 et c_2 sont des nombres quelconques non simultanément nuls.

En particulier, en choisissant d'abord $c_1 = 1$; $c_2 = 0$ et puis $c_1 = 0$; $c_2 = 1$, on obtient le système fondamental des solutions le plus simple composé de deux vecteurs propres de A linéairement indépendants:

$$x^{(1)} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad \text{et} \quad x^{(2)} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

Tous les autres vecteurs propres de A, associés au nombre caractéristique $\lambda_1 = 1$, sont des combinaisons linéaires de ces vecteurs de base et couvrent le plan engendré par les vecteurs $x^{(1)}$ et $x^{(2)}$ (l'origine des coordonnées exceptée).

Prenons maintenant $\lambda_3 = 4$. Portant cette valeur dans l'équation (7) on a:

$$\begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

ou

$$\left.\begin{aligned} -2x_1 + x_2 + x_3 &= 0, \\ x_1 - 2x_2 + x_3 &= 0, \\ x_1 + x_2 - 2x_3 &= 0. \end{aligned}\right\} \qquad (9)$$

Le rang de la matrice (9) est $r = 2$, le mineur supérieur gauche étant

$$\delta = \begin{vmatrix} -2 & 1 \\ 1 & -2 \end{vmatrix} \neq 0.$$

Par suite, la troisième équation du système se déduit des deux premières, et l'on peut se borner au système de deux premières équations

$$\left.\begin{aligned} -2x_1 + x_2 + x_3 &= 0, \\ x_1 - 2x_2 + x_3 &= 0. \end{aligned}\right\}$$

Il en résulte

$$\frac{x_1}{\begin{vmatrix} 1 & 1 \\ -2 & 1 \end{vmatrix}} = \frac{x_2}{-\begin{vmatrix} -2 & 1 \\ 1 & 1 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} -2 & 1 \\ 1 & -2 \end{vmatrix}}$$

ou

$$\frac{x_1}{3} = \frac{x_2}{3} = \frac{x_3}{3}, \quad \text{c'est-à-dire} \quad x_1 = x_2 = x_3 = c,$$

avec c une constante différente de zéro.

En posant $c = 1$ on obtient la solution la plus simple qui réalise le vecteur propre de la matrice A:

$$x^{(3)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

D é f i n i t i o n 2. On dit que le sous-espace linéaire E_k ($k \leqslant n$) est *invariant* par rapport à la transformation linéaire donnée

$$y = Ax,$$

si tout vecteur transformé de ce sous-espace appartient à ce dernier, c'est-à-dire $x \in E_k$ entraîne $Ax \in E_k$.

Il est clair que la démonstration du théorème 1 reste valide si l'on considère la transformation linéaire déterminée par la matrice A dans un espace invariant.

T h é o r è m e 1'. *Toute transformation linéaire déterminée dans un sous-espace invariant d'un espace vectoriel complexe possède au moins un vecteur propre.*

Indiquons encore une propriété importante des vecteurs propres.

T h é o r è m e 2. *Les vecteurs propres d'une matrice, associés aux valeurs propres deux à deux distinctes, sont linéairement indépendants.*

D é m o n s t r a t i o n. Soit A la matrice donnée et

$$Ax^{(j)} = \lambda_j x^{(j)} \quad (j = 1, 2, \ldots, m), \tag{10}$$

où

$$x^{(j)} \neq 0 \quad \text{et} \quad \lambda_j \neq \lambda_k \quad \text{pour } j \neq k.$$

Supposons que

$$c_1 x^{(1)} + c_2 x^{(2)} + \ldots + c_m x^{(m)} = 0, \tag{11}$$

où $|c_1| + |c_2| + \ldots + |c_m| \neq 0$.

Prenons, pour fixer les idées, $c_1 \neq 0$. En appliquant à (11) la transformation A, on a en vertu des formules (10):

$$\lambda_1 c_1 x^{(1)} + \lambda_2 c_2 x^{(2)} + \ldots + \lambda_m c_m x^{(m)} = 0. \tag{12}$$

On en tire en multipliant l'égalité (11) par λ_m et en soustrayant de l'égalité obtenue l'égalité (12):

$$(\lambda_m - \lambda_1) c_1 x^{(1)} + (\lambda_m - \lambda_2) c_2 x^{(2)} + \ldots$$
$$\ldots + (\lambda_m - \lambda_{m-1}) c_{m-1} x^{(m-1)} = 0. \tag{13}$$

Ensuite, d'une façon analogue, on peut éliminer de l'égalité (13) le vecteur $x^{(m-1)}$, etc. En éliminant les vecteurs

$$x^{(m)}, \ x^{(m-1)}, \ \ldots, \ x^{(2)},$$

on obtient

$$(\lambda_m - \lambda_1)(\lambda_{m-1} - \lambda_1) \ldots (\lambda_2 - \lambda_1) c_1 x^{(1)} = 0. \tag{14}$$

Mais cette dernière égalité est impossible, car aucun des facteurs du premier membre n'est égal à zéro. Par conséquent, notre hypothèse est fausse et les vecteurs propres $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ sont linéairement indépendants.

C o r o l l a i r e. Si toutes les valeurs propres de la matrice A d'ordre n sont deux à deux distinctes, les n vecteurs propres * de cette matrice qui leur sont associés forment une base de l'espace correspondant de dimension n.

§ 13. Matrices semblables

D é f i n i t i o n. Deux matrices associées à une même transformation linéaire (réduction) dans des bases différentes sont dites *semblables*.

---

* On suppose que pour chaque valeur propre on prend un seul vecteur propre.

Si la matrice A est semblable à la matrice B, on note $A \backsim B$. Déduisons la condition de similitude de deux matrices. Supposons que la matrice A réalise dans une certaine base la transformation linéaire

$$y = Ax. \tag{1}$$

Dans une nouvelle base (en coordonnées nouvelles) cette même réduction sera décrite par une autre matrice B:

$$\eta = B\xi, \tag{2}$$

où

$$B \backsim A.$$

Désignons par S la matrice de passage du nouveau système à l'ancien, soit

$$x = S\xi, \quad y = S\eta, \tag{3}$$

où

$$\det S \neq 0.$$

En portant les formules (3) dans l'équation (1), on obtient:

$$S\eta = AS\xi.$$

Prémultiplions cette dernière égalité par l'inverse S^{-1} pour obtenir

$$\eta = S^{-1}AS\xi. \tag{4}$$

En comparant les formules (4) et (2) on obtient:

$$B = S^{-1}AS. \tag{5}$$

Pour les matrices A et B liées par la relation (5), on dit que B s'obtient de A *par réduction à l'aide de* S. Ainsi on conclut que *deux matrices sont semblables si et seulement si l'une d'elles s'obtient par réduction de l'autre à l'aide d'une certaine matrice régulière.*

On obtient de l'égalité (5) $A = SBS^{-1}$, c'est-à-dire *si la matrice* B *est semblable à* A, *alors inversement la matrice* A *est aussi semblable à* B. Indiquons certaines propriétés de la réduction à l'aide de la matrice S.

1. *Le transformé d'une somme est égal à la somme des transformés*:

$$S^{-1}(A + B)S = S^{-1}AS + S^{-1}BS.$$

2. *Le transformé d'un produit est égal au produit des transformés*:

$$S^{-1}(AB)S = S^{-1}AS \cdot S^{-1}BS.$$

3. *Le transformé de l'inverse d'une matrice est égal à l'inverse du transformé de la matrice*:

$$S^{-1}A^{-1}S = (S^{-1}AS)^{-1}.$$

4. *Le transformé d'une puissance entière d'une matrice (positive ou négative) est égal à la même puissance du transformé de la matrice:*

$$S^{-1}A^nS = (S^{-1}AS)^n.$$

Théorème 1. *Les matrices semblables ont les mêmes polynômes caractéristiques.*

Démonstration. Soit $B \backsim A$. On demande de montrer que

$$\det(A - \lambda E) = \det(B - \lambda E).$$

Comme

$$B = S^{-1}AS \qquad (\det S \neq 0),$$

il vient

$$\det(B - \lambda E) = \det[S^{-1}(A - \lambda E)S] =$$
$$= \det S^{-1} \det(A - \lambda E) \det S = \det(A - \lambda E)\ ^*.$$

Ainsi

$$\det(B - \lambda E) = \det(A - \lambda E).$$

Corollaire 1. Les matrices semblables possèdent les mêmes traces et les mêmes valeurs propres (ainsi que leurs ordres de multiplicité).

Corollaire 2. La propriété du vecteur d'être propre pour la transformation linéaire donnée ne dépend pas du choix de la base.

En effet, soit

$$Ax = \lambda x \qquad (x \neq 0).$$

Si, dans la nouvelle base, le vecteur x est équivalent au vecteur ξ, on a:

$$x = S\xi,$$

S étant la matrice de passage.

Il en résulte que $AS\xi = \lambda S\xi$ et, par conséquent, $S^{-1}AS\xi = \lambda\xi$, c'est-à-dire ξ est un vecteur propre de la matrice $B = S^{-1}AS \backsim A$ qui décrit dans la nouvelle base notre transformation linéaire.

Remarque. Le polynôme caractéristique, les valeurs propres et les vecteurs propres étant les mêmes pour toutes les matrices qui réalisent la transformation linéaire donnée, ils s'appellent respectivement *polynôme caractéristique*, *valeurs propres et vecteurs propres de la transformation linéaire elle-même.*

Théorème 2. *Si la matrice carrée d'ordre n possède n vecteurs propres linéairement indépendants, en admettant que ces derniers sont de base on obtient une matrice diagonale semblable à la matrice donnée.*

---

* Nous avons appliqué ici les théorèmes connus (cf. chapitre VII, § 2 et § 4): 1) le déterminant du produit de deux matrices carrées du même ordre est égal au produit des déterminants de ces matrices; 2) le déterminant de la matrice inverse est égal à l'inverse du déterminant de la matrice initiale.

D é m o n s t r a t i o n. Soit la matrice carrée A. Formons de ses vecteurs propres $e_1, e_2, \ldots, e_n$ une base. Les vecteurs e_j étant propres, alors

$$Ae_j = \lambda_j e_j \quad (j = 1, 2, \ldots, n).$$

Considérons un vecteur quelconque x de notre espace. En le développant suivant les vecteurs de base e_j $(j = 1, 2, \ldots, n)$, on aura:

$$x = \sum_{j=1}^{\infty} x_j e_j,$$

où x_j sont les coordonnées du vecteur x dans la base donnée.

En rapportant l'application A au vecteur x, on obtient un nouveau vecteur

$$y = Ax = A \sum_{j=1}^{n} x_j e_j$$

ou, la transformation étant linéaire,

$$y = \sum_{j=1}^{n} x_j A e_j = \sum_{j=1}^{n} x_j \lambda_j e_j.$$

Il s'ensuit que les coordonnées du vecteur y dans la base donnée sont

$$y_j = \lambda_j x_j \quad (j = 1, 2, \ldots, n)$$

ou

$$y_j = \sum_{k=1}^{n} \delta_{jk} \lambda_j x_k,$$

où δ_{jk} est le symbole de Kronecker.

Donc, dans la nouvelle base la matrice de la transformation est une matrice diagonale

$$\Lambda = (\delta_{jk} \lambda_j)$$

ou, sous une forme développée,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & \lambda_n \end{bmatrix}.$$

C o r o l l a i r e. Toute matrice carrée, dont les valeurs propres sont deux à deux distinctes, peut être ramenée par similitude à la matrice diagonale.

Ce résultat se déduit immédiatement du théorème 2 du paragraphe précédent.

§ 14. Forme bilinéaire d'une matrice

Soient $A = [a_{jk}]$ une matrice carrée réelle et x, y les vecteurs d'un espace complexe de dimension n. Composons le produit scalaire

$$(Ax, y) = \sum_{j=1}^{n} (Ax)_j y_j^* = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_k y_j^*. \tag{1}$$

L'expression (1) s'appelle *forme bilinéaire* de la matrice A.

Déduisons une propriété importante de la forme bilinéaire. Si l'on modifie l'ordre de sommation tout en changeant entre elles les notations des indices, il est clair que la somme (1) aura sa valeur antérieure. On aboutit donc à

$$(Ax, y) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{kj} x_j y_k^*.$$

Mettons cette somme sous forme de produit scalaire

$$(Ax, y) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{kj} x_j y_k^* = \Big(\sum_{j=1}^{n} \sum_{k=1}^{n} a_{kj} y_k x_j^* \Big)^* =$$
$$= (A'y, x)^* = (x, A'y).$$

Ainsi

$$(Ax, y) = (x, A'y), \tag{2}$$

c'est-à-dire *dans un produit scalaire* (1) *la matrice réelle A peut être portée de la première place à la deuxième en lui substituant sa transposée.*

C o r o l l a i r e. Si la matrice A est une matrice réelle et symétrique ($A' = A$), alors

$$(Ax, y) = (x, Ay), \tag{3}$$

c'est-à-dire *dans un produit scalaire une matrice réelle symétrique peut être portée de la première place à la deuxième.*

§ 15. Propriétés des matrices symétriques

T h é o r è m e 1. *Toute valeur propre d'une matrice symétrique a éléments réels est réelle.*

D é m o n s t r a t i o n. Soient λ une valeur propre de la matrice A et x un vecteur propre correspondant:

$$Ax = \lambda x \qquad (x \neq 0). \tag{1}$$

Comme $A' = A$,

$$(Ax, x) = (x, Ax)$$

ou, en vertu de l'égalité (1),

$$(\lambda x, x) = (x, \lambda x).$$

D'où

$$\lambda (x, x) = \lambda^* (x, x).$$

Un vecteur propre est non nul par définition ; donc

$$(x, x) \neq 0$$

et $\lambda = \lambda^*$, c'est-à-dire λ est un nombre réel.

C o r o l l a i r e. Pour une matrice symétrique réelle, l'équation caractéristique ne possède que des racines réelles.

T h é o r è m e 2. *Les vecteurs propres d'une matrice symétrique réelle, associés à des valeurs propres distinctes, sont orthogonaux entre eux.*

D é m o n s t r a t i o n. Soit A une matrice symétrique réelle. Considérons deux vecteurs propres $x^{(i)}$ et $x^{(j)}$ associés aux valeurs propres λ_i et λ_j ($\lambda_i \neq \lambda_j$). On a :

$$Ax^{(i)} = \lambda_i x^{(i)} \tag{2}$$

et

$$Ax^{(j)} = \lambda_j x^{(j)}. \tag{3}$$

Composons le produit scalaire

$$(Ax^{(i)}, x^{(j)}) = (x^{(i)}, Ax^{(j)}).$$

En vertu des égalités (2) et (3) on a :

$$(\lambda_i x^{(i)}, x^{(j)}) = (x^{(i)}, \lambda_j x^{(j)})$$

et

$$\lambda_i (x^{(i)}, x^{(j)}) = \lambda_j^* (x^{(i)}, x^{(j)}). \tag{4}$$

La valeur propre λ_j étant réelle en vertu du théorème 1, $\lambda_j^* = \lambda_j$. La formule (4) entraîne donc

$$(\lambda_i - \lambda_j) (x^{(i)}, x^{(j)}) = 0.$$

Or

$$\lambda_i - \lambda_j \neq 0$$

et

$$(x^{(i)}, x^{(j)}) = 0,$$

c'est-à-dire que les vecteurs propres $x^{(i)}$ et $x^{(j)}$ sont orthogonaux entre eux.

R e m a r q u e. On peut admettre que les vecteurs propres d'une matrice symétrique aux éléments réels sont réels.

Théorème 3. *Toute matrice symétrique réelle peut être ramenée par réduction à une matrice diagonale.*

Démonstration. Pour rendre la démonstration immédiate bornons-nous au cas de l'espace E_3 de dimension trois.

Soit une matrice symétrique A d'ordre trois. On sait que toute matrice a au moins un vecteur propre (§ 12, théorème 1). Désignons par e_1 le vecteur propre de A. Cette matrice étant symétrique, on peut choisir le vecteur e_1 réel.

Considérons tous les vecteurs x orthogonaux au vecteur e_1, c'est-à-dire tels que

$$(x, e_1) = 0. \tag{5}$$

Montrons que ces vecteurs forment un sous-espace invariant E_2 par rapport à la transformation A (fig. 55).

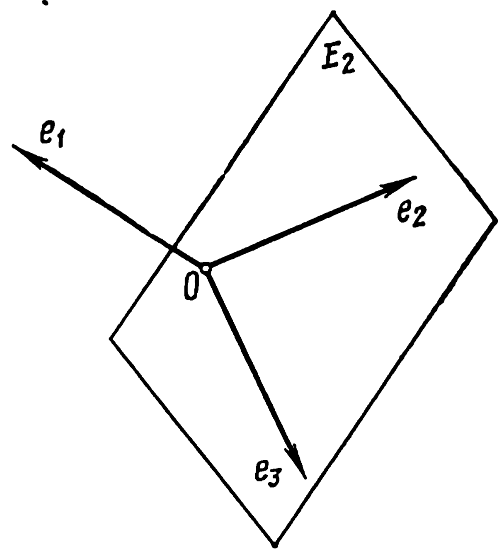


Fig. 55.

En effet, si $x \in E_2$ et $y \in E_2$, c'est-à-dire si

$$(x, e_1) = (y, e_1) = 0,$$

on a pour tout α et β:

$$(\alpha x + \beta y, e_1) =$$
$$= \alpha (x, e_1) + \beta (y, e_1) = 0$$

et, par conséquent,

$$\alpha x + \beta y \in E_2.$$

Ainsi, l'ensemble des vecteurs qui vérifient la condition (5) forme un espace linéaire et on voit aisément que c'est un espace de dimension deux.

Soit maintenant $x \in E_2$. Considérons le produit scalaire

$$(Ax, e_1) = (x, Ae_1) = (x, \lambda_1 e_1) = \lambda_1 (x, e_1) = 0,$$

c'est-à-dire

$$Ax \in E_2.$$

En vertu du théorème 1′ (§ 12), dans un espace E_2 de dimension deux, il existe également un vecteur propre e_2 de la matrice A. Considérons maintenant les vecteurs x orthogonaux au vecteur e_1 ainsi qu'au vecteur e_2, c'est-à-dire tels que

$$(x, e_1) = (x, e_2) = 0.$$

On montre d'une façon analogue que ces vecteurs forment un espace E_1 de dimension un invariant par rapport à la transformation A. L'espace E_1 possède également un vecteur propre e_3 de la matrice

A. Les vecteurs e_1, e_2, e_3, étant orthogonaux deux à deux, sont linéairement indépendants. Ainsi, on construit la base orthogonale de l'espace E_3 composée des vecteurs propres de la matrice A.

Désignons par λ_j les valeurs propres associées aux vecteurs propres e_j. En vertu du théorème 2 du § 13, la matrice Λ de la transformation linéaire donnée, rapportée à la base propre e_1, e_2, e_3, sera une matrice diagonale; de plus, dans notre cas

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}.$$

D'une façon analogue on démontre le théorème pour le cas général.

C o r o l l a i r e 1. Pour toute transformation linéaire à matrice symétrique réelle, il existe une base orthogonale composée des vecteurs propres réels de la matrice donnée, dans laquelle la matrice de la transformation est diagonale.

C o r o l l a i r e 2. Si la matrice est symétrique, les vecteurs propres linéairement indépendants associés à chacune de ses valeurs propres sont comptés autant de fois que l'ordre de multiplicité de cette valeur propre l'indique.

T h é o r è m e 4 (p r o p r i é t é e x t r é m a l e d e s v a l e u r s p r o p r e s). *Soit A une matrice symétrique réelle et*

$$\lambda = \min(\lambda_1, \lambda_2, \ldots, \lambda_n),$$
$$\Lambda = \max(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

où $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont toutes les valeurs propres de A.

Alors, l'inégalité

$$\lambda (x, x) \leqslant (Ax, x) \leqslant \Lambda (x, x) \tag{6}$$

est vérifiée pour tout vecteur x.

D é m o n s t r a t i o n. En vertu du corollaire 1 du théorème 3 la matrice A possède un système des vecteurs propres e_1, e_2, ..., e_n

$$Ae_j = \lambda_j e_j \quad (j = 1, 2, \ldots, n),$$

qui forment une base orthonormale de l'espace E_n. Alors tout vecteur x peut être mis sous la forme

$$x = x_1 e_1 + x_2 e_2 + \ldots + x_n e_n,$$

$x_1, x_2, \ldots, x_n$ étant les coordonnées du vecteur x rapportées à la base donnée. Par suite

$$Ax = x_1 Ae_1 + x_2 Ae_2 + \ldots + x_n Ae_n = \lambda_1 x_1 e_1 + \lambda_2 x_2 e_2 + \ldots + \lambda_n x_n e_n.$$

En tenant compte de ce que les vecteurs de la base sont orthogo-

naux, on aura:

$$(Ax, x)=\left(\sum_{j=1}^{n}\lambda_j x_j e_j, \sum_{k=1}^{n} x_k e_k\right)=\sum_{j=1}^{n}\sum_{k=1}^{n}\lambda_j x_j x_k^* (e_j, e_k)=$$

$$=\sum_{j=1}^{n}\sum_{k=1}^{n}\lambda_j x_j x_k^*\delta_{jk}=\sum_{j=1}^{n}\lambda_j |x_j|^2,$$

c'est-à-dire

$$(Ax, x)=\sum_{j=1}^{n}\lambda_j |x_j|^2. \tag{7}$$

En remplaçant dans l'égalité (7) λ_j par la plus petite valeur de λ, on obtient:

$$(Ax, x)\geqslant\lambda\sum_{j=1}^{n}|x_j|^2=\lambda(x, x).$$

De façon analogue, en substituant à λ_j dans l'égalité (7) la valeur maximale Λ, on trouve:

$$(Ax, x)\leqslant\Lambda\sum_{j=1}^{n}|x_j|^2=\Lambda(x, x).$$

Ainsi, l'inégalité (6) est démontrée.

C o r o l l a i r e. Les valeurs propres minimale λ et maximale Λ d'une matrice symétrique réelle A sont respectivement les valeurs minimale et maximale de la forme quadratique

$$u=(Ax, x)$$

sur la sphère unité $(x, x)=1$.

En effet, en posant dans l'inégalité (6) $(x, x)=1$, on aura:

$$\lambda\leqslant(Ax, x)\leqslant\Lambda.$$

De plus, si $Ax=\lambda x$, il vient

$$(Ax, x)=(\lambda x, x)=\lambda;$$

d'une façon analogue, si $Ax=\Lambda x$, alors

$$(Ax, x)=(\Lambda x, x)=\Lambda.$$

Ainsi,

$$\lambda=\min(Ax, x)\quad\text{pour}\quad(x, x)=1$$

et

$$\Lambda=\max(Ax, x)\quad\text{pour}\quad(x, x)=1.$$

La matrice symétrique réelle $A=[a_{ij}]$ est appelée matrice *définie positive* si la forme quadratique correspondante

$$u=(Ax, x)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_i x_j^*$$

est définie positive (cf. chapitre VIII, § 13), c'est-à-dire que pour tout vecteur $x \neq 0$ on a:

$$(Ax, x) > 0.$$

T h é o r è m e 5. *Une matrice symétrique réelle est définie positive si et seulement si toutes ses valeurs propres sont positives.*

D é m o n s t r a t i o n. Si A une matrice symétrique réelle et ses valeurs propres λ_j sont telles que $\lambda_j > 0$ $(j = 1, 2, \ldots, n)$, la formule (7) de la démonstration du théorème précédent amène:

$$(Ax, x) = \sum_{j=1}^{n} \lambda_j |x_j|^2,$$

où $x = (x_1, x_2, \ldots, x_n)$. D'où pour $x \neq 0$

$$(Ax, x) > 0,$$

et la matrice A est définie positive.

Inversement, soit A une matrice symétrique réelle définie positive. En vertu du théorème 1, toutes ses valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont réelles et

$$\lambda = \min(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

est la plus petite valeur de la forme quadratique $u = (Ax, x)$ sur la sphère $(x, x) = 1$. La forme quadratique étant positive sur cette sphère, on a donc:

$$\lambda > 0.$$

On en tire à plus forte raison

$$\lambda_j > 0 \quad \text{pour} \quad j = 1, 2, \ldots, n.$$

Voici sans démonstration les conditions de définition positive d'une matrice réelle [2].

T h é o r è m e 6. *Pour qu'une matrice réelle* $A = [a_{ij}]$, *avec* $a_{ij} = a_{ji}$, *soit définie positive, il faut et il suffit que les conditions de Sylvester*

$$\Delta_1 = a_{11} > 0; \quad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0; \ \ldots;$$

$$\Delta_n = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} > 0$$

soient remplies, c'est-à-dire la matrice symétrique réelle A est définie positive si et seulement si les mineurs diagonaux principaux de son déterminant det A sont strictement positifs.

§ 16*. Propriétés des matrices à éléments réels

Dans ce qui suit nous allons étudier essentiellement les matrices $A = [a_{ij}]$ dont les éléments a_{ij} sont réels.

Soit $A = [a_{ij}]$ une matrice carrée réelle d'ordre n. Son équation caractéristique

$$\det (A - \lambda E) = 0$$

étant un polynôme aux coefficients réels, si les racines $\lambda_1, \lambda_2, \ldots, \lambda_n$ de l'équation caractéristique, qui représentent les valeurs propres de A, sont complexes, elles sont conjuguées deux à deux (chapitre V, § 1), c'est-à-dire si λ_s est une valeur propre de A, le nombre conjugué λ_s^* est également une valeur propre de A de même ordre de multiplicité.

Il se peut qu'une matrice réelle ne possède pas de valeurs propres réelles. Toutefois, dans un cas particulier important, lorsque les éléments d'une matrice sont positifs, on assure l'existence au moins d'une valeur propre réelle [6].

Théorème de Perron. *Si tous les éléments d'une matrice carrée sont positifs, sa valeur propre la plus grande en module est également positive et est une racine simple de l'équation caractéristique de la matrice ; à cette racine est associée un vecteur propre de coordonnées positives.*

Les vecteurs propres d'une matrice réelle A à valeurs propres distinctes sont dans le cas général complexes et ne jouissent pas de la propriété d'orthogonalité. Cependant, en faisant appel aux vecteurs propres de la transposée A', on peut obtenir ce qu'on appelle les *relations de biorthogonalité* qui, pour le cas d'une matrice symétrique, sont équivalentes aux relations d'orthogonalité ordinaires.

Théorème 1. *Si la matrice A est réelle et ses valeurs propres sont deux à deux distinctes, il existe deux bases $\{x_j\}$ et $\{x'_j\}$ de l'espace E_n composées respectivement de vecteurs propres de A et de vecteurs propres de la transposée A', qui vérifient les conditions biorthonormales suivantes:*

$$(x_j, x'_k) = \begin{cases} 0 & \text{pour } j \neq k, \\ 1 & \text{pour } j = k. \end{cases}$$

Démonstration. Soient $\lambda_1, \lambda_2, \ldots, \lambda_n$ les valeurs propres de la matrice A. Etant donné que la matrice A est réelle, ses valeurs propres sont conjuguées deux à deux, c'est-à-dire que si λ_j est sa valeur propre, le nombre conjugué λ_j^* l'est également. Désignons par x_j $(j = 1, 2, \ldots, n)$ les vecteurs propres correspondants de A

$$Ax_j = \lambda_j x_j \quad (j = 1, 2, \ldots, n). \tag{1}$$

Les vecteurs $\{x_j\}$ forment une base de l'espace E_n (§ 12, théorème 2, corollaire).

Comme le déterminant ne change pas sa valeur lors du remplacement des lignes par des colonnes,

$$\det(A' - \lambda E) \equiv \det(A - \lambda E)$$

et, par conséquent, la transposée A' de la matrice a les mêmes valeurs propres λ_j que A. Soient x'_j $(j = 1, 2, \ldots, n)$ les vecteurs propres de la matrice A' associés aux valeurs propres conjuguées λ_j^*:

$$A'x'_j = \lambda_j^* x'_j \quad (j = 1, 2, \ldots, n). \tag{2}$$

Les vecteurs $\{x'_j\}$ forment également une base de l'espace E_n.

Les bases $\{x_j\}$ et $\{x'_j\}$ sont *biorthogonales*:

$$(x_j, x'_k) = 0 \quad \text{pour} \quad j \neq k. \tag{3}$$

En effet, d'une part on a:

$$(Ax_j, x'_k) = (\lambda_j x_j, x'_k) = \lambda_j (x_j, x'_k). \tag{4}$$

D'autre part, compte tenu du fait que la matrice A est réelle, on obtient:

$$(Ax_j, x'_k) = (x_j, A'x'_k) = (x_j, \lambda_k^* x'_k) = \lambda_k (x_j, x'_k). \tag{5}$$

Les égalités (4) et (5) entraînent

$$\lambda_j (x_j, x'_k) = \lambda_k (x_j, x'_k). \tag{6}$$

Comme $\lambda_j \neq \lambda_k$ pour $j \neq k$, l'égalité (6) entraîne l'égalité (3). Montrons que les vecteurs $\{x_j\}$ et $\{x'_j\}$ peuvent être normés de façon que

$$(x_j, x'_j) = 1 \quad (j = 1, 2, \ldots, n). \tag{7}$$

En effet, développant le vecteur x_j par rapport aux vecteurs de la base $\{x'_1, x'_2, \ldots, x'_n\}$, on aura:

$$x_j = c_1 x'_1 + \ldots + c_j x'_j + \ldots + c_n x'_n.$$

D'où, compte tenu de la condition de biorthogonalité (3),

$$(x_j, x_j) = c_1^* (x_j, x'_1) + \ldots + c_j^* (x_j, x'_j) + \ldots$$
$$\ldots + c_n^* (x_j, x'_n) = c_j^* (x_j, x'_j) > 0;$$

et

$$(x_j, x'_j) = \alpha_j \neq 0.$$

En prenant au lieu des vecteurs $x'_1, \ldots, x'_n$ les vecteurs $\frac{1}{\alpha_1^*} x'_1, \ldots$ $\ldots, \frac{1}{\alpha_n^*} x'_n$, on obtient la norme (7) nécessaire du fait que

$$\left(x_j, \frac{1}{\alpha_j^*} x'_j\right) = \frac{1}{\alpha_j} (x_j, x'_j) = \frac{1}{\alpha_j} \cdot \alpha_j = 1 \quad (j = 1, 2, \ldots, n).$$

Ainsi, si les valeurs propres d'une matrice réelle A sont distinctes, on peut toujours trouver pour une base propre $\{x_j\}$ de A une base propre $\{x'_j\}$ de la transposée A' telle que

$$(x_j, x'_k) = \delta_{jk}, \tag{8}$$

où δ_{jk} est le symbole de Kronecker.

C o r o l l a i r e. Si la matrice A est réelle et symétrique $(A' = A)$, on peut poser: $x'_j = x_j$ $(j = 1, 2, \ldots, n)$, où x_j sont les vecteurs propres normés de A (cf. § 15). Il vient

$$(x_j, x_k) = \delta_{jk}.$$

Déduisons encore un *développement bilinéaire d'une matrice A*.

T h é o r è m e 2. *Soient A une matrice réelle carrée et*

$$X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}$$

$(j = 1, 2, \ldots, n)$ ses vecteurs propres considérés comme matrices colonnes et

$$X'_k = [x'_{1k} \ldots x'_{nk}]$$

*$(k = 1, 2, \ldots, n)$ les vecteurs propres respectifs * de la transposée A considérés comme matrices lignes, les conditions de biorthonormalité* (8)

$$(X_j, X'^{*}_k) = X'_k X_j = \delta_{jk} \tag{9}$$

étant vérifiées.

Alors, on a la relation

$$A = \lambda_1 X_1 X'_1 + \lambda_2 X_2 X'_2 + \ldots + \lambda_n X_n X'_n, \tag{10}$$

où $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont les valeurs propres de la matrice A.

D é m o n s t r a t i o n. Considérons les matrices

$$X = \begin{bmatrix} x_{11} & \cdot & x_{1n} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdot & x_{nn} \end{bmatrix} \quad \text{et} \quad X' = \begin{bmatrix} x'_{11} & \ldots & x'_{n1} \\ \cdot & \cdot\,\cdot\,\cdot\,\cdot & \cdot \\ x'_{1n} & \ldots & x'_{nn} \end{bmatrix},$$

composées respectivement des colonnes X_j $(j = 1, \ldots, n)$ et des lignes X'_k $(k = 1, \ldots, n)$.

L'égalité (9) entraîne

$$X'X = [\sum_{k=1}^{n} x'_{ki} x_{kj}] = [X_j X'_i] = [\delta_{ji}] = E, \tag{11}$$

* C'est-à-dire associés aux mêmes valeurs propres des matrices A et A'.

où E est la matrice unité. Comme la matrice X est composée de colonnes linéairement indépendantes, elle est régulière, det $X \neq 0$, et il existe donc l'inverse X^{-1}. En vertu de l'égalité (11) (cf. chapitre VII, § 4, théorème, remarque 1), on a:

$$X^{-1} = X'.$$

On en déduit

$$XX' = E,$$

et on obtient ainsi les *deuxièmes relations de biorthogonalité* [7]

$$\sum_{k=1}^{n} x_{ik}x'_{jk} = \delta_{ij}. \qquad (12)$$

En appliquant ces relations, on a:

$$X_1X'_1 + X_2X'_2 + \ldots + X_nX'_n = [x_{i1}x'_{j1}] +$$
$$+ [x_{i2}x'_{j2}] + \ldots + [x_{in}x'_{jn}] = [\sum_{k=1}^{n} x_{ik}x'_{jk}] = [\delta_{ij}] = E,$$

c est-à-dire

$$E = X_1X'_1 + X_2X'_2 + \ldots + X_nX'_n.$$

En multipliant à gauche cette égalité par A et compte tenu de

$$AX_j = \lambda_j X_j \quad (j = 1, 2, \ldots, n),$$

on aboutit évidemment à l'égalité (10).

Notons que dans l'égalité (10) X_j et X'_j $(j = 1, 2, \ldots, n)$ sont des vecteurs propres de A et A', associés à la même valeur propre λ_j malgré les notations du théorème 1, où x_j et x'_j sont des vecteurs propres de A et A', associés aux valeurs propres λ_j et λ_j^* complexes conjuguées.

BIBLIOGRAPHIE

G. Chilov. Introduction à la théorie des espaces linéaires. Gostekhizdat. Moscou-Léningrad, 1952, chapitres I-IX.
I. Guelfand. Cours d'algèbre linéaire, éd. 2. Gostekhizdat, Moscou-Léningrad, 1951, chapitres I-II.
A. Maltsev. Principes d'algèbre linéaire. Gostekhizdat, Moscou-Léningrad, 1948, chapitres I-III.
A. S. Housholder. Principles of Numerical Analysis. Mc. Graw-Hill, 1953, chapitre II.
I. Schreider. Résolution des systèmes d'équations linéaires algébriques. Comptes rendus de l'Académie des Sciences de l'U.R.S.S., 5, 1951.
F. Gantmacher. Théorie des matrices. Gostekhizdat, Moscou, 1953, chapitre VIII.
V .Faddéeva. Méthodes numériques de l'algèbre linéaire. Gostekhizdat, Moscou-Léningrad, 1950, chapitre I.

CHAPITRE XI*

SUPPLÉMENTS SUR LA CONVERGENCE DES PROCESSUS ITÉRATIFS DES SYSTÈMES D'ÉQUATIONS LINÉAIRES

§ 1. Convergence des séries matricielles entières

Théorème 1. *Une série matricielle entière*

$$\sum_{k=0}^{\infty} a_k X^k \tag{1}$$

à coefficients numériques a_k converge si toutes les valeurs propres λ_1, $\lambda_2, \ldots, \lambda_n$ de la matrice X se situent dans le cercle de convergence fermé $|x| \leqslant R$ (fig. 56) *de la série scalaire entière*

$$\sum_{k=0}^{\infty} a_k x^k \tag{2}$$

($x = \xi + i\eta$), *les valeurs propres reposant sur la circonférence du cercle de convergence étant simples et constituant des points de convergence de la série* (2).

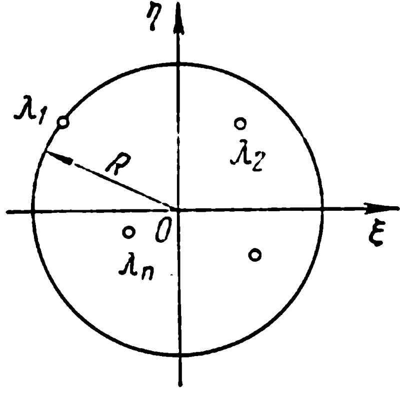


Fig. 56.

Une série (1) *diverge si au moins une valeur propre de X se trouve en dehors du cercle de convergence fermé de la série* (2) *ou s'il existe une valeur propre de X reposant sur la circonférence du cercle de convergence pour lequel la série* (2) *diverge.*

Démonstration. 1) Soit la matrice X telle que

$$|\lambda_1| \leqslant R, \ldots, |\lambda_n| \leqslant R.$$

Supposons pour simplifier que les valeurs propres λ_j ($j = 1, 2, \ldots, n$) de X soient simples. La matrice X peut alors être diagonalisée à l'aide d'une matrice régulière S

$$X = S^{-1} [\lambda_1, \ldots, \lambda_n] S.$$

Introduisons les notations

$$F_m(X) = \sum_{k=0}^{m} a_k X^k, \quad f_m(x) = \sum_{k=0}^{m} a_k x^k$$

et

$$f(x)=\lim_{m\to\infty} f_m(x)=\sum_{k=0}^{\infty} a_k x^k.$$

On a

$$F_m(X)=\sum_{k=0}^{m} a_k \{S^{-1}[\lambda_1, \ldots, \lambda_n] S\}^k = S^{-1}\left\{\sum_{k=0}^{m} a_k [\lambda_1^k, \ldots, \lambda_n^k]\right\} S =$$
$$= S^{-1}[f_m(\lambda_1), \ldots, f_n(\lambda_n)] S. \quad (3)$$

Puisque les nombres λ_j se trouvent à l'intérieur du cercle de convergence de la série entière (2) ou bien coïncident avec les points de convergence de cette série, lesdits points appartenant à la circonférence du cercle de convergence, il existe des limites finies

$$f(\lambda_j)=\lim_{m\to\infty} f_m(\lambda_j) \quad (j=1\ 2, \ldots, n).$$

En passant dans la formule (3) à la limite quand $m\to\infty$ on amène:

$$F(X)=\lim_{m\to\infty} F_m(X)=S^{-1}[f(\lambda_1), \ldots f(\lambda_n)] S, \quad (4)$$

c'est-à-dire la série matricielle (1) converge en X.

On peut démontrer que le théorème est vrai également pour des valeurs propres multiples λ_j, mais nous n'examinerons pas ce cas [1].

2) Supposons, par exemple, qu'au moins une valeur propre λ_1 de la matrice X soit telle que

$$|\lambda_1| > R.$$

Comme λ_1 repose hors du cercle de convergence de la série entière (2), quand $m\to\infty$, $f_m(\lambda_1)$ n'a pas de limite. La formule (3) entraîne que, lorsque $m\to\infty$, $F_m(X)$ n'a pas non plus de limite, c'est-à-dire la série (1) diverge en X.

Un résultat analogue s'obtient si $|\lambda_1|=R$ et la série $\sum_{k=0}^{\infty} a_k\lambda_1^k$ est divergente.

Remarque. D'après (4), si $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont des valeurs propres simples de X, alors $f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)$, où

$$f(x)=\sum_{k=0}^{\infty} a_k x^k,$$

sont les valeurs propres de la fonction

$$F(X)=\sum_{k=0}^{\infty} a_k X^k.$$

En particulier, les valeurs propres de la matrice X^k sont les nombres $\lambda_1^k, \ldots, \lambda_n^k$.

Théorème 2. *La progression géométrique matricielle*

$$E + X + X^2 + \ldots + X^k + \ldots, \tag{5}$$

où X est une matrice carrée d'ordre n, converge si et seulement si toutes les valeurs propres

$$\lambda_j = \lambda_j(X) \qquad (j = 1, 2, \ldots, n)$$

de X reposent à l'intérieur du cercle unité

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n); \tag{6}$$

de plus, si la série (5) *est divergente*, $X^k \not\to 0$ pour $k \to \infty$.

Démonstration. En effet, puisque pour la série entière correspondante

$$\sum_{k=0}^{\infty} x^k \tag{7}$$

le rayon de convergence $R = 1$, pour $|x| = 1$ la série (7) étant divergente, en vertu du théorème 1, la progression géométrique (5) ne converge que si les conditions (6) sont remplies.

Si la série (5) est divergente, alors

$$|\lambda_j| \geqslant 1 \qquad (j = 1, 2, \ldots, n).$$

D'où on a, en supposant pour simplifier que les valeurs propres $\lambda_1, \ldots, \lambda_n$ sont distinctes,

$$X = S^{-1}[\lambda_1, \ldots, \lambda_n] S,$$

S étant une matrice régulière. Donc

$$X^k = S^{-1}[\lambda_1^k, \ldots, \lambda_n^k] S,$$

et, par conséquent, $X^k \not\to 0$ pour $k \to \infty$. Cette dernière affirmation est vraie aussi pour des valeurs multiples de λ_j, mais nous ne nous attarderons pas sur ce fait.

Théorème 3. *Toute valeur propre $\lambda_1, \ldots, \lambda_n$ d'une matrice carrée X ne dépasse en module aucune de ses normes canoniques*

$$|\lambda_j| \leqslant \| X \| \quad (j = 1, 2, \ldots, n).$$

Démonstration. Posons

$$\| X \| = \rho$$

et considérons la matrice

$$Y = \frac{1}{\rho + \varepsilon} X, \tag{8}$$

avec $\varepsilon > 0$. Il est clair que

$$\| Y \| = \frac{1}{\rho+\varepsilon} \| X \| = \frac{\rho}{\rho+\varepsilon} < 1.$$

Par suite (chapitre VII, § 10, théorème 5), la série

$$E + Y + Y^2 + \ldots + Y^k + \ldots$$

converge.

On en déduit en vertu du théorème (2) que les valeurs propres $\mu_1, \ldots, \mu_n$ de la matrice Y verifient les inégalités

$$| \mu_j | < 1 \qquad (j = 1, 2, \ldots, n).$$

Mais la formule (8) entraîne

$$\mu_j = \frac{1}{\rho+\varepsilon} \lambda_j \qquad (j = 1, 2, \ldots, n).$$

Donc

$$| \lambda_j | < \rho + \varepsilon \qquad (j = 1, 2, \ldots, n)$$

ou, vu que le nombre ε est arbitraire,

$$| \lambda_j | \leqslant \rho = \| X \| \qquad (j = 1, 2, \ldots, n),$$

ce qu'il fallait démontrer.

§ 2. Identité d'Hamilton-Cayley

Théorème. *Toute matrice carrée X est une racine de son polynôme caractéristique, c'est-à-dire si*

$$\psi(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n,$$

où $\psi(\lambda) = \det(\lambda E - X)$, *alors*

$$\psi(X) = X^n + p_1 X^{n-1} + \ldots + p_n E \equiv 0.$$

Démonstration. Supposons que toutes les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ de X, c'est-à-dire les racines de l'équation caractéristique $\psi(\lambda) = 0$, soient distinctes. La matrice X peut être alors diagonalisée à l'aide d'une matrice régulière S:

$$X = S^{-1} [\lambda_1, \lambda_2, \ldots, \lambda_n] S.$$

Comme $\psi(X)$ est un cas particulier de la série matricielle entière, la formule (4) du § 1 entraîne

$$\psi(X) = S^{-1} [\psi(\lambda_1), \psi(\lambda_2), \ldots, \psi(\lambda_n)] S.$$

Mais il est évident que

$$\psi(\lambda_j) = 0 \qquad (j = 1, 2, \ldots, n).$$

Il vient donc

$$\psi(X) = S^{-1} [0, 0, \ldots, 0] S = 0.$$

Si l'équation caractéristique $\psi(\lambda) = 0$ possède des racines multiples, elles peuvent être considérées comme les limites des racines distinctes, lorsqu'on donne aux coefficients de l'équation des écarts infiniment petits [1]. Il en résulte que le théorème se généralise à ce cas aussi.

§ 3. Conditions nécessaires et suffisantes de la convergence du processus itératif d'un système linéaire

En utilisant les valeurs propres d'une matrice $\alpha = [\alpha_{ij}]$ on peut donner les conditions nécessaires et suffisantes de la convergence du processus itératif d'un système linéaire

$$x = \alpha x + \beta, \tag{1}$$

avec

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad \text{et} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Théorème. *Pour que le processus itératif*

$$x^k = \alpha x^{(k-1)} + \beta$$
$$(k = 1, 2, \ldots) \tag{2}$$

converge quel que soit le choix du vecteur initial $x^{(0)}$ et quel que soit le terme constant β, il faut et il suffit que les valeurs propres de α, c'est-à-dire les racines de l'équation caractéristique

$$\det(\alpha - \lambda E) = 0$$

soient en module inférieures à un.

Démonstration. La formule (2) entraîne:

$$x^k = (E + \alpha + \alpha^2 + \ldots + \alpha^{k-1})\beta + \alpha^k x^{(0)}. \tag{3}$$

On en déduit que la convergence du processus itératif (2), β et $x^{(0)}$ étant arbitraires, est équivalente à la convergence de la progression géométrique matricielle

$$E + \alpha + \alpha^2 + \ldots = \sum_{k=0}^{\infty} \alpha^k. \tag{4}$$

En vertu du théorème 2 du § 1 la progression géométrique (4) converge si toutes les valeurs propres λ_j $(j = 1, 2, \ldots, n)$ de α vérifient les inégalités

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n). \tag{5}$$

Puisque dans ces conditions $\alpha^k \to 0$ quand $k \to \infty$, la formule (3) conduit à un processus itératif convergent quels que soient β et $x^{(0)}$,

c'est-à-dire il existe une limite

$$\lim_{k\to\infty} x^k = x,$$

où x est évidemment une solution du système (1).

Si les inégalités (5) ne sont pas observées, la série (4) diverge. Dans ce cas, pour un certain choix du terme constant et du vecteur initial $x^{(0)}$, le processus itératif diverge également.

Ainsi, pour rendre convergent le processus itératif (2), il faut et il suffit que toutes les racines $\lambda_1, \lambda_2, \ldots, \lambda_n$ de l'équation caractéristique

$$\begin{vmatrix} \alpha_{11}-\lambda & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22}-\lambda & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn}-\lambda \end{vmatrix} = 0$$

vérifient les conditions $|\lambda_j| < 1$ $(j = 1, 2, \ldots, n)$.

Corollaire. Pour que le processus itératif (2) converge il suffit que

$$\|\alpha\| < 1, \tag{6}$$

pour une norme canonique (cf. chapitre IX, § 1).

En effet, en vertu du théorème 3 du § 1, l'inégalité (6) conduit à l'inégalité (5).

Remarque. Considérons le système linéaire

$$Ax = b, \tag{7}$$

où $A = [a_{ij}]$ et $b = [b_1, \ldots b_n]$ est un vecteur colonne.

Soit

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix} \neq 0.$$

Pour réduire le système (7) au type spécial (1) on pose en général :

$$A = D + (A - D).$$

D'où

$$Dx = b - (A - D)\, x,$$

et puisque $\det D = a_{11}a_{22} \ldots a_{nn} \neq 0$,

$$x = D^{-1}b + D^{-1}(D - A)\, x.$$

On peut adopter

$$\alpha = D^{-1}(D - A).$$

Ainsi pour que le processus itératif ordinaire du système linéaire (7) converge quels que soient le terme constant b et le vecteur initial $x^{(0)}$, il faut et il suffit que toutes les racines $\lambda_1, \lambda_2, \ldots, \lambda_n$ de l'équation caractéristique

$$\det [D^{-1}(D - A) - \lambda E] = 0 \tag{8}$$

soient en module inférieures à un. Si l'on applique le théorème sur le déterminant du produit de deux matrices, l'équation (8) peut être transformée de la façon suivante:

$$\det D^{-1} \det [(D - A) - \lambda D] = 0$$

ou

$$\det [\lambda D + (A - D)] = 0,$$

c'est-à-dire

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}\lambda & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn}\lambda \end{vmatrix} = 0.$$

§ 4. Conditions nécessaires et suffisantes de la convergence du processus de Seidel pour un système linéaire

Considérons pour le système linéaire

$$x = \alpha x + \beta, \tag{1}$$

où $\alpha = [\alpha_{ij}]$ et $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$, le processus de Seidel

$$x_i^{(k)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k-1)} + \beta_i \quad (i = 1, 2, \ldots, n;\ k = 1, 2, \ldots)$$

le vecteur initial arbitraire étant

$$x^{(0)} = \begin{bmatrix} x_1^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}.$$

Posons

$$\alpha = B + C,$$

où

$$B = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ \alpha_{21} & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{n,n-1} & 0 \end{bmatrix}, \quad C = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ 0 & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \alpha_{nn} \end{bmatrix}.$$

Le processus de Seidel peut se mettre alors sous la forme matricielle suivante

$$x^{(k)} = Bx^{(k)} + Cx^{(k-1)} + \beta \quad (k = 1, 2, \ldots). \tag{2}$$

Théorème. *Pour que le processus de Seidel* (2) *du système* (1) *soit convergent quel que soit le choix du terme constant* β *et du vecteur initial* $x^{(0)}$, *il faut et il suffit que toutes les racines* $\lambda_1, \ldots, \lambda_n$ *de l'équation*

$$\det [C - (E - B)\lambda] \equiv \begin{vmatrix} \alpha_{11} - \lambda & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21}\lambda & \alpha_{22} - \lambda & \ldots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1}\lambda & \alpha_{n2}\lambda & \ldots & \alpha_{nn} - \lambda \end{vmatrix} = 0 \tag{3}$$

soient inférieures en module à un.

Démonstration. Il résulte de la formule (2)

$$(E - B)\, x^{(k)} = Cx^{(k-1)} + \beta. \tag{4}$$

La matrice $E - B$ est régulière du fait que

$$\det (E - B) = 1.$$

Aussi, l'égalité (4) peut s'écrire

$$x^{(k)} = (E - B)^{-1} Cx^{(k-1)} + (E - B)^{-1} \beta. \tag{5}$$

Il est évident donc que le processus de Seidel est équivalent à une simple itération appliquée au système linéaire

$$x = (E - B)^{-1} Cx + (E - B)^{-1} \beta.$$

En vertu du théorème du paragraphe précédent, pour que le processus (5) soit convergent il faut et il suffit que les racines $\lambda_1, \ldots, \lambda_n$ de l'équation caractéristique

$$\det [(E - B)^{-1} C - \lambda E] = 0 \tag{6}$$

vérifient les conditions

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n).$$

L'équation (6) est évidemment équivalente à l'équation (3).

Remarque. Soit

$$Ax = b. \tag{7}$$

Posons

$$D = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix} \neq 0.$$

Pour appliquer la méthode de Seïdel, dans les cas courants le système (7) s'écrit

$$Dx = (D - A)\,x + b$$

ou

$$x = D^{-1}(D - A)\,x + D^{-1}b. \tag{8}$$

Posons

$$A - D = B_1 + C_1,$$

avec

$$B_1 = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{n,\,n-1} & 0 \end{bmatrix}$$

et

$$C_1 = \begin{bmatrix} 0 & a_{12} & \dots & a_{1,\,n-1} & a_{1n} \\ 0 & 0 & \dots & a_{2,\,n-1} & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Alors, il vient

$$D^{-1}(D - A) = B + C,$$

où

$$B = -D^{-1}B_1 \text{ et } C = -D^{-1}C_1,$$

de plus, les matrices triangulaires B et C réalisent la partition de la matrice du système (8) nécessaire pour appliquer le processus de Seidel. D'après la formule (3), la convergence du processus de Seidel pour le système (7) est définie par les propriétés des racines de l'équation

$$\det\,[-D^{-1}C_1 - (E + D^{-1}B_1)\,\lambda] = 0. \tag{9}$$

L'équation (9) peut être remplacée par une équation équivalente

$$\det\,[(D + B_1)\,\lambda + C_1] = 0$$

ou

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1}\lambda & a_{n2}\lambda & \dots & a_{nn}\lambda \end{vmatrix} = 0. \tag{10}$$

Ainsi pour que le processus de Seidel appliqué au système (7) soit convergent, quels que soient le terme constant b et l'approximation initiale $x^{(0)}$, il faut et il suffit que toutes les racines λ_j de l'équation (10) satisfassent aux conditions

$$|\lambda_j| < 1 \qquad (j = 1, 2, \dots, n).$$

§ 5. Convergence du processus de Seidel pour un système normal

Théorème. *Pour un système normal le processus de Seidel converge quel que soit le choix du vecteur initial.*

Démonstration. Supposons que le système linéaire

$$Ax = b \tag{1}$$

soit normal, c'est-à-dire que la matrice $A = [a_{ij}]$ soit symétrique et définie positive.

Adoptons

$$A = D + V + V^*,$$

où

$$D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

est une matrice diagonale,

$$V = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}$$

une matrice triangulaire inférieure, et

$$V^* = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

une matrice triangulaire supérieure, transposée de V par suite de la forme symétrique de A. On a alors:

$$(D + V + V^*)\,x = b.$$

D'où

$$Dx = b - (V + V^*)\,x$$

et

$$x = D^{-1}b - D^{-1}(V + V^*)\,x, \tag{2}$$

avec

$$D^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & \dots & 0 \\ 0 & 0 & \dots & \frac{1}{a_{nn}} \end{bmatrix}.$$

D'après ce qui précède, le processus de Seidel du système (1) ou du système (2), équivalent à (1), se construit de la façon suivante:

$$x^{(k)} = D^{-1}b + Bx^{(k)} + Cx^{(k-1)} \quad (k = 1, 2, \ldots), \tag{3}$$

avec

$$B = -D^{-1}V \text{ et } C = -D^{-1}V^*.$$

En vertu du théorème du paragraphe précédent, pour que le processus converge, il faut et il suffit que toutes les valeurs propres λ de la matrice

$$M = (E - B)^{-1}C$$

soient en module inférieures à un.

Dans notre cas on a:

$$M = -(E + D^{-1}V)^{-1} D^{-1}V^* = -[D^{-1}(D + V)]^{-1} D^{-1}V^* =$$
$$= -(D + V)^{-1} DD^{-1}V^* = -(D + V)^{-1} V^*.$$

Soit e un vecteur unité propre de la matrice M associé à la valeur propre λ

$$(D+V)^{-1} V^*e = -\lambda e$$

ou

$$V^*e = -\lambda(D+V)e.$$

On en tire

$$(V^*e, e) = -\lambda[(D+V)e, e]$$

et

$$\lambda = -\frac{(V^*e, e)}{(De, e) + (Ve, e)}.$$

Introduisons les notations

$$(De, e) = \sum_{j=1}^{n} a_{jj} |e_j|^2 = \sigma > 0$$

et

$$(Ve, e) = \alpha + i\beta$$

(α et β réels et $i^2 = -1$).

La matrice V^* étant la transposée de V, on obtient:

$$(V^*e, e) = (e, Ve) = (Ve, e)^* = \alpha - i\beta.$$

Donc

$$\lambda = -\frac{\alpha - i\beta}{(\sigma + \alpha) + i\beta}$$

et donc

$$|\lambda| = \frac{\sqrt{\alpha^2 + \beta^2}}{\sqrt{(\sigma + \alpha)^2 + \beta^2}}. \tag{4}$$

Le fait que A est définie positive donne

$$(Ae, e) = (De, e) + (Ve, e) + (V^*e, e) = \\ = \sigma + (\alpha + i\beta) + (\alpha - i\beta) = \sigma + 2\alpha > 0,$$

c'est-à-dire

$$\sigma + \alpha > -\alpha. \tag{5}$$

Ensuite, la positivité du nombre σ conduit évidemment à:

$$\sigma + \alpha > \alpha.$$

Ainsi, l'inégalité

$$\sigma + \alpha > |\alpha| \tag{6}$$

est toujours vraie. Il en résulte pour les termes de la fraction (4):

$$\sqrt{(\sigma+\alpha)^2+\beta^2} > \sqrt{\alpha^2+\beta^2} \geqslant 0,$$

ou

$$|\lambda| < 1,$$

ce qu'il fallait démontrer.

§ 6. Vérification efficace des conditions de convergence

Pour vérifier les conditions du théorème de convergence des processus itératifs, il faut disposer de critères efficaces permettant de définir si les racines $\lambda_1, \lambda_2, \ldots, \lambda_n$ du polynôme algébrique donné

$$f(\lambda) = p_0\lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

vérifient ou ne vérifient pas la condition

$$|\lambda_j| < 1 \qquad (j = 1, 2, \ldots, n). \tag{2}$$

Ce problème peut être résolu simplement en faisant appel aux *conditions de Hurwicz* connues [2].

Théorème de Hurwicz. *Supposons que les coefficients p_k $(k = 0, 1, \ldots, n)$ du polynôme* (1) *soient réels, que $p_0 > 0$ et que*

$$M = \begin{bmatrix} p_1 & p_0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ p_3 & p_2 & p_1 & p_0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & p_n & p_{n-1} & p_{n-2} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & p_n \end{bmatrix}$$

soit une matrice d'ordre n dont les lignes sont des suites des coefficients du polynôme (1)

$$p_{2m-1}, p_{2m-2}, \ldots, p_{2m-n},$$

où on a posé $p_k = 0$ pour $k < 0$ et $k > n$. Alors, toutes les racines

$\lambda_1, \lambda_2, \ldots, \lambda_n$ de (1) possèdent des parties réelles négatives

$$\operatorname{Re}\lambda_j < 0 \qquad (j = 1, 2, \ldots, n)$$

(c'est-à-dire se situent dans le demi-plan gauche du plan complexe $\lambda = \alpha + i\beta$), si et seulement si les mineurs diagonaux principaux de M sont positifs, soit

$$\left.\begin{array}{l} \Delta_1 = p_1 > 0, \\ \Delta_2 = \begin{vmatrix} p_1 & p_0 \\ p_3 & p_2 \end{vmatrix} > 0, \\ \ldots\ldots\ldots\ldots \\ \Delta_n = p_n\Delta_{n-1} > 0. \end{array}\right\} \qquad (3)$$

Exemple 1. Pour le trinôme du deuxième degré

$$p_0\lambda^2 + p_1\lambda + p_2$$

les conditions de Hurwicz s'écrivent

$$p_0 > 0, \quad p_1 > 0, \quad p_2 > 0.$$

Nous voulons déterminer le cas où les racines du polynôme (1) vérifient la condition (2), c'est-à-dire reposent dans le plan complexe λ à l'intérieur du cercle unité

$$|\lambda| < 1.$$

La fonction homographique

$$\lambda = \frac{\mu+1}{\mu-1}$$

permet de transformer l'intérieur du cercle unité $|\lambda| < 1$ en un demi-plan gauche $\operatorname{Re}\mu < 0$. Le polynôme (1) se met alors sous la forme

$$f\left(\frac{\mu+1}{\mu-1}\right) = p_0\left(\frac{\mu+1}{\mu-1}\right)^n + p_1\left(\frac{\mu+1}{\mu-1}\right)^{n-1} + \ldots + p_n =$$
$$= \frac{1}{(\mu-1)^n}\left[p_0(\mu+1)^n + p_1(\mu+1)^{n-1}(\mu-1) + \ldots + p_n(\mu-1)^n\right].$$

Il en résulte que les racines de (1) reposent à l'intérieur du cercle unité si et seulement si le polynôme auxiliaire

$$F(\mu) = \pm\left[p_0(\mu+1)^n + p_1(\mu+1)^{n-1}(\mu-1) + \ldots \right.$$
$$\left. \ldots + p_n(\mu-1)^n\right]$$

satisfait aux conditions de Hurwicz (3), le signe étant choisi de façon que le coefficient du terme principal

$$\pm(p_0 + p_1 + \ldots + p_n) > 0.$$

Exemple 2. Considérons le trinôme du deuxième degré

$$f(\lambda) = \lambda^2 + p\lambda + q, \qquad (4)$$

où p et q sont réels. Le polynôme auxiliaire est de la forme

$$F(\mu) = \pm [(\mu + 1)^2 + p(\mu + 1)(\mu - 1) + q(\mu - 1)^2] =$$
$$= \pm [(1 + p + q)\mu^2 + 2(1 - q)\mu + (1 - p + q)].$$

Les conditions de Hurwicz donnent

$$\left.\begin{array}{l} \pm(1+p+q) > 0, \\ \pm(1-q) > 0, \\ \pm(1-p+q) > 0. \end{array}\right\}$$

Considérons les cas:

a) $q < 1$, alors $q > -p - 1$ et $q > p - 1$;

b) $q > 1$, alors $q < -p - 1$ et $q < p - 1$,

ce qui est impossible.

Donc les racines λ_1, λ_2 de l'équation (4) sont inférieures en module à un si et seulement si

$$|p| < 1 + q, \quad |q| < 1. \tag{5}$$

Puisque pour $n = 2$, l'équation caractéristique de la matrice α s'écrit

$$\begin{vmatrix} \alpha_{11} - \lambda & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - \lambda \end{vmatrix} = 0,$$

ou

$$\lambda^2 - (\alpha_{11} + \alpha_{22})\lambda + \det \alpha = 0,$$

alors, pour que le processus itératif correspondant d'un système de deux équations converge, il faut que

$$|\det \alpha| < 1.$$

En général, les domaines de convergence d'un processus itératif ordinaire et d'un processus de Seidel se superposent. Il existe des systèmes linéaires tels que le processus itératif ordinaire converge alors que le processus de Seidel soit divergent et inversement [3].

E x e m p l e 3. Considérons le système linéaire

$$x = \alpha x + \beta \tag{6}$$

à matrice antisymétrique

$$\alpha = \begin{bmatrix} p & q \\ -q & p \end{bmatrix},$$

p et q étant réels.

L'équation caractéristique de α est de la forme

$$\begin{vmatrix} p - \lambda & q \\ -q & p - \lambda \end{vmatrix} = 0$$

ou

$$(\lambda - p)^2 + q^2 = 0.$$

D'où

$$\lambda_{1,2} = p \pm iq.$$

Pour que la méthode itérative ordinaire converge, il faut et il suffit que

$$|\lambda_{1,2}| = \sqrt{p^2 + q^2} < 1,$$

c'est-à-dire

$$p^2 + q^2 < 1$$

(domaine A de la figure 57).

Pour la méthode de Seidel l'équation qui définit la convergence s'écrit

$$\begin{vmatrix} p-\lambda & q \\ -q\lambda & p-\lambda \end{vmatrix} = 0$$

ou

$$\lambda^2 - (2p - q^2)\,\lambda + p^2 = 0. \quad (7)$$

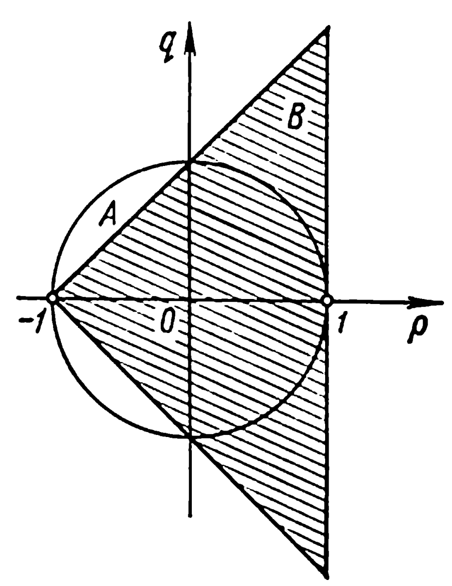


Fig. 57.

En vertu des résultats de l'exemple (2), pour que les racines λ_1 et λ_2 de l'équation (7) vérifient les conditions

$$|\lambda_1| < 1, \qquad |\lambda_2| < 1,$$

il faut et il suffit de respecter les inégalités

$$|2p - q^2| < 1 + p^2, \qquad p^2 < 1,$$

d'où

$$|p| < 1, \qquad |q| < 1 + p$$

(domaine B de la figure 57). Les domaines A et B se superposent partiellement; il s'ensuit que pour le système (6) on peut choisir les coefficients p et q premièrement tels que la méthode itérative converge et que la méthode de Seidel diverge (par exemple, $p = -0{,}5$; $q = 0{,}6$), et deuxièmement, tels que la méthode de Seidel converge et que la méthode itérative diverge (par exemple $p = 0{,}5$; $q = 1$).

BIBLIOGRAPHIE

1. *V. Smirnov.* Cours de mathématiques supérieures, t. III. Editions Mir, Moscou, 1970.
2. *A. Kurosh.* Cours d'algèbre supérieure. Editions Mir, Moscou, 1971.
3. *V. Faddéeva.* Méthodes numériques de l'algèbre linéaire. Gostekhizdat, Moscou-Léningrad, 1950, chapitre II.

CHAPITRE XII

CALCUL DES VALEURS PROPRES ET DES VECTEURS PROPRES D'UNE MATRICE

§ 1. Notes d'introduction

Il arrive souvent que pour résoudre des problèmes théoriques et pratiques il faille déterminer les valeurs propres de la matrice A donnée, c'est-à-dire calculer les racines de son équation *caractéristique* (*séculaire*)

$$\det (A - \lambda E) = 0, \tag{1}$$

ainsi que ses vecteurs propres associés. Le deuxième problème est plus simple, car si les racines de l'équation caractéristique sont connues, la recherche des vecteurs propres se ramène à l'obtention de solutions non nulles de certains systèmes linéaires homogènes. Nous allons donc en premier lieu étudier le calcul des racines de l'équation caractéristique (1).

A cet effet on fait surtout appel à deux procédés: 1) développement du déterminant caractéristique en un polynôme de degré n

$$D(\lambda) = \det (A - \lambda E)$$

et résolution de l'équation $D(\lambda) = 0$ par l'un des procédés approchés connus (par exemple, par la méthode de Lobatchevski-Graeffe, cf. chapitre V, §§ 7-12) et 2) approximation des racines de l'équation caractéristique (le plus souvent maximales en module) par la méthode itérative sans développer au préalable le déterminant caractéristique.

Nous exposerons dans ce chapitre les méthodes principales de résolution du problème général énoncé, en commençant par le *développement des déterminants caractéristiques.*

§ 2. Développement des déterminants caractéristiques

Ainsi qu'il est connu, le *déterminant caractéristique* d'une matrice $A = [a_{ij}]$ est un déterminant du type

$$D(\lambda) = \det (A - \lambda E) = \begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix}. \tag{1}$$

En l'annulant on obtient une *équation caractéristique*

$$D(\lambda) = 0.$$

Si le problème consiste à trouver toutes les racines de cette équation, il convient de calculer d'abord le déterminant (1).

Le développement de (1) donne un polynôme de degré n

$$D(\lambda) = (-1)^n [\lambda^n - \sigma_1\lambda^{n-1} + \sigma_2\lambda^{n-2} - \ldots + (-1)^n \sigma_n], \qquad (2)$$

où

$$\sigma_1 = \sum_{\alpha=1}^{n} a_{\alpha\alpha}$$

est la somme de tous les mineurs diagonaux d'ordre un de la matrice A;

$$\sigma_2 = \sum_{\alpha<\beta} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta} \end{vmatrix}$$

est la somme de tous les mineurs diagonaux d'ordre deux de la matrice A;

$$\sigma_3 = \sum_{\alpha<\beta<\gamma} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} & a_{\alpha\gamma} \\ a_{\beta\alpha} & a_{\beta\beta} & a_{\beta\gamma} \\ a_{\gamma\alpha} & a_{\gamma\beta} & a_{\gamma\gamma} \end{vmatrix}$$

est la somme de tous les mineurs diagonaux d'ordre trois de la matrice A, etc. Enfin

$$\sigma_n = \det A.$$

On voit aisément que le nombre de mineurs diagonaux d'ordre k de la matrice A s'écrit

$$C_n^k = \frac{n(n-1)\ldots(n-k+1)}{k!} \qquad (k = 1, 2, \ldots, n).$$

On en tire que le calcul immédiat des coefficients du polynôme caractéristique (2) est équivalent au calcul de

$$C_n^1 + C_n^2 + \ldots + C_n^n = 2^n - 1$$

déterminants de divers ordres. Ce dernier problème est difficile à réaliser dès que les valeurs de n deviennent quelque peu grandes. Aussi a-t-on conçu à cet effet des méthodes spéciales (méthodes de Krylov, de Danilevski, de Leverrier, méthodes des coefficients indéterminés, d'interpolation, etc.) (cf. [1]). Dans les paragraphes qui suivent nous exposerons certaines de ces méthodes.

§ 3. Méthode de Danilevski

Cette méthode consiste en principe à ramener le déterminant caractéristique à la *forme normale de Frobénius*

$$D(\lambda)=\begin{vmatrix} p_1-\lambda & p_2 & p_3 & \dots & p_n \\ 1 & -\lambda & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\lambda \end{vmatrix}. \qquad (1)$$

Si nous parvenons à mettre le déterminant caractéristique sous la forme (1), on obtient en le développant suivant la première ligne

$$D(\lambda)=(p_1-\lambda)(-\lambda)^{n-1}-p_2(-\lambda)^{n-2}+p_3(-\lambda)^{n-3}-\dots+(-1)^{n-1}p_n$$

ou

$$D(\lambda)=(-1)^n(\lambda^n-p_1\lambda^{n-1}-p_2\lambda^{n-2}-p_3\lambda^{n-3}-\dots-p_n). \qquad (2)$$

Ainsi, développer le déterminant mis sous la forme (1) ne présente aucune difficulté. Désignons par

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

la matrice donnée et par

$$P=\begin{bmatrix} p_1 & p_2 & \dots & p_{n-1} & p_n \\ 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}$$

une *matrice de Frobénius*, semblable à la première, c'est-à-dire

$$P=S^{-1}AS,$$

S étant une matrice régulière.

Les polynômes caractéristiques des matrices semblables étant les mêmes, on a :

$$\det(A-\lambda E)=\det(P-\lambda E). \qquad (3)$$

Pour justifier la méthode il suffit donc de montrer comment on construit P à partir de A. D'après la méthode de Danilevski, pour passer de la matrice A à la matrice semblable P, on effectue $n-1$ réductions qui transforment successivement les lignes de la matrice A à partir de la dernière, en lignes respectives de la matrice P.

Montrons le départ du processus. Il nous faut réduire la ligne

$$a_{n1}a_{n2}\ldots a_{n,\,n-1}a_{nn}$$

pour obtenir la ligne 0 0 ... 1 0. En supposant que $a_{n,\,n-1}\neq 0$, divisons tous les éléments de la $(n-1)$-ième colonne de A par $a_{n,\,n-1}$. La n-ième ligne s'écrit alors

$$a_{n1}a_{n2}\ldots a_{nn}.$$

Retranchons ensuite la $(n-1)$-ième colonne de la matrice transformée multipliée respectivement par les nombres $a_{n1}, a_{n2}, \ldots, a_{nn}$ de toutes ses autres colonnes.

Il en résulte une matrice dont la dernière ligne a la forme cherchée 0 0 ... 1 0. Les opérations indiquées sont des transformations élémentaires appliquées aux colonnes de la matrice A. En appliquant ces mêmes transformations à la matrice unité on obtient la matrice

$$M_{n-1}=\begin{bmatrix} 1 & 0 & \ldots & 0 & 0\\ 0 & 1 & \ldots & 0 & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ m_{n-1,\,1} & m_{n-1,\,2} & \ldots & m_{n-1,\,n-1} & m_{n-1,\,n}\\ 0 & 0 & \ldots & 0 & 1\end{bmatrix},$$

où

$$m_{n-1,\,i}=-\frac{a_{ni}}{a_{n,\,n-1}}\quad\text{avec}\quad i\neq n-1 \tag{4}$$

et

$$m_{n-1,\,n-1}=\frac{1}{a_{n,\,n-1}}. \tag{4'}$$

On en tire (cf. chapitre VII, § 14) que les opérations effectuées sont équivalentes à la prémultiplication de la matrice M_{n-1} par la matrice A, c'est-à-dire après les transformations indiquées on obtient la matrice

$$AM_{n-1}=B=\begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1,\,n-1} & b_{1,\,n}\\ b_{21} & b_{22} & \ldots & b_{2,\,n-1} & b_{2,\,n}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ b_{n-1,\,1} & b_{n-1,\,2} & \ldots & b_{n-1,\,n-1} & b_{n-1,\,n}\\ 0 & 0 & \ldots & 1 & 0\end{bmatrix}. \tag{5}$$

L'application de la règle de multiplication des matrices donne les formules pour calculer les éléments de la matrice B:

$$b_{ij}=a_{ij}+a_{i,\,n-1}m_{n-1,\,j}\quad\text{avec}\quad 1\leqslant i\leqslant n;\ j\neq n-1; \tag{6}$$

$$b_{j,\,n-1}=a_{i,\,n-1}m_{n-1,\,n-1}\quad\text{avec}\quad 1\leqslant i\leqslant n. \tag{6'}$$

Toutefois, la matrice $B = AM_{n-1}$ ne sera pas semblable à la matrice A. Pour réaliser une réduction, il faut postmultiplier l'inverse M_{n-1}^{-1} par la matrice B:

$$M_{n-1}^{-1}AM_{n-1} = M_{n-1}^{-1}B.$$

La vérification directe montre facilement que l'inverse M_{n-1}^{-1} est de la forme

$$M_{n-1}^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ a_{n1} & a_{n2} & \dots & a_{n,\,n-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}. \qquad (7)$$

Soit

$$M_{n-1}^{-1}AM_{n-1} = C.$$

Donc

$$C = M_{n-1}^{-1}B. \qquad (8)$$

Puisqu'il est évident que la postmultiplication de M_{n-1}^{-1} par B ne change pas la ligne transformée de cette dernière, la matrice C s'écrit

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1,\,n-1} & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2,\,n-1} & c_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n-1,\,1} & c_{n-1,\,2} & \dots & c_{n-1,\,n-1} & c_{n-1,\,n} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \qquad (9)$$

En multipliant les matrices M_{n-1}^{-1} (7) et B (5), on a:

$$c_{ij} = b_{ij} \quad \text{avec} \quad 1 \leqslant i \leqslant n-2 \qquad (10)$$

et

$$c_{n-1,\,j} = \sum_{k=1}^{n} a_{nk}b_{kj} \quad \text{avec} \quad 1 \leqslant j \leqslant n. \qquad (10')$$

Ainsi la multiplication de M_{n-1}^{-1} par B ne change que la $(n-1)$-ième ligne de B. Les éléments de cette ligne s'obtiennent d'après les formules (10) et (10′). La matrice C obtenue est semblable à la matrice A et possède une ligne réduite. La première étape du processus est ainsi achevée.

Ensuite, si $c_{n-1,\,n-2} \neq 0$, on peut refaire les opérations analogues sur la matrice C en prenant comme ligne du pivot sa $(n-2)$-ième ligne. Il en résulte la matrice

$$D = M_{n-2}^{-1}CM_{n-2}$$

à deux lignes réduites. Soumettons cette matrice aux mêmes opérations. En poursuivant ce processus on obtient finalement la matrice

de Frobénius:

$$P=M_1^{-1}\dots M_{n-2}^{-1}M_{n-1}^{-1}AM_{n-1}M_{n-2}\dots M_1,$$

si, certes, toutes les $n-1$ transformations intermédiaires sont possibles.

Tout ce processus peut être traduit par un schéma de calcul commode dont la composition est illustrée par l'exemple suivant.

E x e m p l e. Ramener à la forme de Frobénius la matrice

$$A=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}.$$

S o l u t i o n. Portons les résultats du calcul sur le tableau 25.

Inscrivons sur les 1-4-ièmes lignes du tableau les éléments a_{ij} $(i,\ j=1,\ 2,\ 3,\ 4)$ de la matrice donnée et les sommes de contrôle $a_{i5}=\sum_{j=1}^{4}a_{ij}$ $(i=1,\ 2,\ 3,\ 4)$ (Σ). Marquons l'élément $a_{43}=2$ figurant dans la troisième colonne (*colonne marquée*). Portons sur la ligne I les éléments de la troisième ligne de la matrice $M_{n-1}=M_3$ calculés d'après les formules (4) et (4′):

$$m_{31}=-\frac{a_{41}}{a_{43}}=-\frac{4}{2}=-2;$$

$$m_{32}=-\frac{a_{42}}{a_{43}}=-\frac{3}{2}=-1{,}5;$$

$$m_{33}=\frac{1}{a_{43}}=\frac{1}{2}=0{,}5;$$

$$m_{34}=-\frac{a_{44}}{a_{43}}=-\frac{1}{2}=0{,}5.$$

On place sur cette même ligne (I) l'élément

$$m_{35}=-\frac{a_{45}}{a_{43}}=-\frac{10}{2}=-5,$$

obtenu d'une façon analogue à partir de la colonne de contrôle Σ. Le nombre -5 doit coïncider avec la somme des éléments de la ligne I qui ne font pas partie de la colonne de contrôle (après le remplacement de l'élément m_{33} par -1). Par commodité, inscrivons le nombre -1 à côté de l'élément m_{33} en les séparant par un trait.

Inscrivons sur les lignes 5 à 8 de la colonne M^{-1} la troisième ligne de la matrice M^{-1} qui, en vertu de la formule (7), coïncide avec la quatrième ligne de la matrice initiale A. Portons sur les lignes 5 à 8

Schéma de calcul de Danilevski

Numéro de la ligne	M^{-1}	Colonnes de la matrice 1	2	3	4	Σ	Σ′
1		1	2	3	4	10	
2		2	1	2	3	8	
3		3	2	1	2	8	
4		4	3	2	1	10	
1	M_3^{-1} | M_3	−2	−1,5	0,5 | −1	−0,5	−5	
5	4	−5	−2,5	1,5	2,5	−3,5	−5
6	3	2	−2	1	2	−1	−2
7	2	1	0,5	0,5	1,5	3,5	3
8	1	0	0	1	0	1	0
7′		−24	−15	11	19	−9	
11	M_2^{-1} | M_2	−1,600	−0,067 | −1	0,733	1,267	−0,600	
9	−24	−1	0,167	−0,333	−0,667	−1,833	−2
10	−15	1,2	0,133	−0,467	−0,533	−0,333	0,2
11	11	0	1	0	0	1	0
12	19	0	0	1	0	1	1
10′		6	5	34	24	69	
111	M_1^{-1} | M_1	0,167 | −1	−0,833	−5,667	−4,000	−11,500	
13	6	−0,167	1	5,333	3,333	9,500	9,667
14	5	1	0	0	0	1	0
15	34	0	1	0	0	1	1
16	24	0	0	1	0	1	1
13′		4	40	56	20	120	

et sur les colonnes correspondantes les éléments de la matrice

$$B = AM_3,$$

calculés d'après les formules à deux termes (6) pour les colonnes non marquées et d'après la formule à un terme (6') pour la colonne marquée. Par exemple, pour la première colonne on a:

$$b_{11} = 1 + 3(-2) = -5;$$
$$b_{21} = 2 + 2(-2) = -2;$$
$$b_{31} = 3 + 1(-2) = 1;$$
$$b_{41} = 4 + 2(-2) = 0;$$

etc.

Les éléments transformés de la troisième colonne (marquée) s'obtiennent en multipliant les éléments initiaux par $m_{33} = 0{,}5$. Par exemple,

$$b_{13} = 3 \cdot 0{,}5 = 1{,}5;$$
$$b_{23} = 2 \cdot 0{,}5 = 1;$$
$$b_{33} = 1 \cdot 0{,}5 = 0{,}5;$$
$$b_{43} = 2 \cdot 0{,}5 = 1.$$

Constatons que la deuxième ligne de la matrice B doit s'écrire

$$0 \ 0 \ 1 \ 0.$$

Pour vérifier, complétons B par les éléments correspondants de la colonne Σ transformés suivant les formules à deux termes analogues, avec $m_{35} = -5$. Par exemple,

$$b_{16} = 10 + 3 \cdot (-5) = -5;$$
$$b_{26} = 8 + 2 \cdot (-5) = -2;$$
$$b_{36} = 8 + 1 \cdot (-5) = 3;$$
$$b_{46} = 10 + 2 \cdot (-5) = 0.$$

Inscrivons les résultats obtenus sur les lignes correspondantes de la colonne Σ'. En leur ajoutant les éléments de la troisième colonne, on obtient les sommes de contrôle

$$b_{i5} = \sum_{j=1}^{4} b_{ij} \qquad (i = 1, 2, 3, 4)$$

pour les lignes 5 à 8 (colonne Σ).

La transformation M_3^{-1} de la matrice B qui donne la matrice $C = M_3^{-1}B$ ne change que la troisième ligne de B, c'est-à-dire la septième ligne de la matrice. Les éléments de cette ligne transformée 7' s'obtiennent d'après la formule (10), c'est-à-dire ce sont des sommes des produits pairs des éléments de la colonne M^{-1}, figurant

sur les lignes 5 à 8, par les éléments correspondants de chacune des colonnes de la matrice B. Par exemple,

$$c_{31} = 4(-5) + 3(-2) + 2 \cdot 1 = -24,$$

etc.

Opérons de même sur la colonne Σ :

$$c_{35} = 4(-3{,}5) + 3(-1) + 2 \cdot 3{,}5 + 1 \cdot 1 = -9.$$

Il en résulte la matrice C composée de lignes 5, 6, 7′, 8 aux sommes de contrôle Σ, la matrice C étant semblable à la matrice A et possédant une ligne réduite 8. Là prend fin la première réduction $C = M_3^{-1}AM_3$.

Ensuite, en prenant la matrice C pour initiale et en prélevant l'élément $c_{32} = -15$ (deuxième colonne), on poursuit la procédure d'une façon analogue. Il en résulte la matrice $D = M_2^{-1}CM_2$ dont les éléments figurent sur les lignes 9, 10′, 11, 12 et qui contient deux lignes réduites. Enfin, en partant de l'élément $d_{21} = 6$ (première colonne) et en transformant la matrice D en une matrice semblable, on obtient la matrice de Frobénius P cherchée dont les éléments figurent sur les lignes 13′, 14, 15, 16. A chaque étape de la procédure, le contrôle se fait à l'aide des colonnes Σ et Σ'.

Ainsi la matrice de Frobénius s'écrit :

$$P = \begin{bmatrix} 4 & 40 & 56 & 20 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

On en déduit que le déterminant caractéristique réduit à la forme normale de Frobénius est de la forme :

$$D(\lambda) = \begin{bmatrix} 4-\lambda & 40 & 56 & 20 \\ 1 & -\lambda & 0 & 0 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & -\lambda \end{bmatrix}$$

ou

$$D(\lambda) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20.$$

§ 4. Cas particuliers de la méthode de Danilevski

Le processus de Danilevski ne présente aucun inconvénient si tout élément marqué est différent du zéro. Nous examinerons dans ce qui suit les cas particuliers lorsque cette restriction n'est pas observée.

Supposons que la transformation de la matrice A en une matrice de Frobénius P aboutit après quelques pas à la matrice

$$D=\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} & \dots & d_{1,\,n-1} & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2k} & \dots & d_{2,\,n-1} & d_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ d_{k1} & d_{k2} & \dots & d_{kk} & \dots & d_{k,\,n-1} & d_{kn} \\ 0 & 0 & \dots & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & \dots & 1 & 0 \end{bmatrix},$$

et il s'avère que $d_{k,\,k-1} = 0$.

La transformation par la méthode de Danilevski devient alors impossible. Deux cas peuvent se présenter.

1. Supposons qu'un élément quelconque de D, à gauche de l'élément nul $d_{k,\,k-1}$, soit différent du zéro, c'est-à-dire $d_{k,\,l} \neq 0$, où $l < k - 1$. Cet élément est alors porté à la place de $d_{k,\,k-1}$, c'est-à-dire nous permutons les $(k-1)$-ième et l-ième colonnes de D en permutant simultanément ses $(k-1)$-ième et l-ième lignes. On peut montrer que la nouvelle matrice D' sera semblable à l'ancienne. Appliquons à la nouvelle matrice la méthode de Danilevski.

2. Soit $d_{kl} = 0$ $(l = 1, 2, \dots, k-1)$; alors D s'écrit

$$D=\left[\begin{array}{cccc|cccc} & (D_1) & & & & (L) & & \\ c_{11} & c_{12} & \dots & c_{1,\,k-1} & c_{1k} & \dots & c_{1,\,n-1} & c_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{k-1,\,1} & c_{k-1,\,2} & \dots & c_{k-1,\,k-1} & c_{k-1,\,k} & \dots & c_{k-1,\,n-1} & c_{k-1,\,n} \\ \hline 0 & 0 & \dots & 0 & c_{kk} & \dots & c_{k,\,n-1} & c_{kn} \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & 0 & \dots & 1 & 0 \\ & & (0) & & & & (D_2) & \end{array}\right] = \left[\begin{array}{c|c} D_1 & L \\ \hline 0 & D_2 \end{array}\right].$$

Dans ce cas le déterminant caractéristique $\det (D - \lambda E)$ se décompose en deux déterminants

$$\det (D - \lambda E) = \det (D_1 - \lambda E) \det (D_2 - \lambda E).$$

La matrice D_2 est déjà ramenée à la forme canonique de Frobénius et de ce fait le calcul de det $(D_2 - \lambda E)$ est immédiat. Il reste à appliquer la méthode de Danilevski à la matrice D_1.

§ 5. Calcul des vecteurs propres par la méthode de Danilevski

Si l'on connaît les valeurs propres d'une matrice, la méthode de Danilevski permet de déterminer ses vecteurs propres. Soit λ une valeur propre de la matrice A et, donc, une valeur propre de la matrice de Frobénius P semblable à A.

Cherchons le vecteur propre $y = (y_1, y_2, \ldots, y_n)$ de P qui correspond à la valeur donnée de λ: $Py = \lambda y$. D'où $(P - \lambda E)\, y = 0$ ou

$$\begin{bmatrix} p_1-\lambda & p_2 & p_3 & \ldots & p_n \\ 1 & -\lambda & 0 & \ldots & 0 \\ 0 & 1 & -\lambda & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & -\lambda \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = 0.$$

En multipliant les matrices entre elles on obtient un système pour déterminer les coordonnées $y_1, y_2, \ldots, y_n$ du vecteur propre y:

$$\left.\begin{aligned} (p_1-\lambda)\, y_1 + p_2 y_2 + \ldots + p_n y_n &= 0, \\ y_1 - \lambda y_2 \qquad\qquad &= 0, \\ y_2 - \lambda y_3 \qquad &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y_{n-1} - \lambda y_n &= 0. \end{aligned}\right\} \qquad (1)$$

Le système (1) est homogène. Ses solutions peuvent s'obtenir de la façon suivante à un coefficient de proportionnalité près. Posons $y_n = 1$. Alors, on a successivement:

$$\left.\begin{aligned} y_{n-1} &= \lambda, \\ y_{n-2} &= \lambda^2, \\ &\ldots\ldots \\ y_1 &= \lambda^{n-1}. \end{aligned}\right\} \qquad (2)$$

Ainsi le vecteur propre cherché est

$$y = \begin{bmatrix} \lambda^{n-1} \\ \lambda^{n-2} \\ \vdots \\ 1 \end{bmatrix}.$$

Désignons maintenant par x le vecteur propre de A associé à la valeur λ. On a évidemment:

$$x = M_{n-1}M_{n-2} \ldots M_2M_1y.$$

La transformation M_1 sur y conduit à

$$M_1y = \begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} m_{1k}y_k \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} m_{1k}y_k \\ \lambda^{n-2} \\ \vdots \\ 1 \end{bmatrix}.$$

La transformation M_1 ne change donc que la première coordonnée du vecteur. Une transformation analogue M_2 ne change que la deuxième coordonnée du vecteur M_1y, etc. En reprenant ce processus $n - 1$ fois, on obtient le vecteur propre x cherché de la matrice A.

§ 6. Méthode de Krylov

Examinons la méthode du développement du déterminant caractéristique due à A. Krylov [2] dont le principe diffère foncièrement de celui de Danilevski. Soit

$$D(\lambda) \equiv \det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

un polynôme caractéristique (à un signe près) de la matrice A. Suivant l'identité de Hamilton-Cayley (chapitre XI, § 2), la matrice A annule son polynôme caractéristique, et donc

$$A^n + p_1A^{n-1} + \ldots + p_nE = 0. \tag{2}$$

Prenons maintenant un vecteur non nul quelconque

$$y^{(0)} = \begin{bmatrix} y_1^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix}.$$

La postmultiplication de deux membres de l'égalité (2) par $y^{(0)}$ donne:

$$A^ny^{(0)} + p_1A^{n-1}y^{(0)} + \ldots + p_ny^{(0)} = 0. \tag{3}$$

Posons

$$A^ky^{(0)} = y^{(k)} \quad (k = 1, 2, \ldots, n); \tag{4}$$

l'égalité (3) se met alors sous la forme

$$y^{(n)} + p_1y^{(n-1)} + \ldots + p_ny^{(0)} = 0 \tag{5}$$

ou

$$\begin{bmatrix} y_1^{(n-1)} & y_1^{(n-2)} \cdot y_1^{(0)} \\ y_2^{(n-1)} & y_2^{(n-2)} \cdot y_2^{(0)} \\ \vdots & \vdots \ \vdots \vdots \\ y_n^{(n-1)} & y_n^{(n-2)} \cdot y_n^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = -\begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix}, \qquad (5')$$

avec

$$y^{(k)} = \begin{bmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \vdots \\ y_n^{(k)} \end{bmatrix} \qquad (k = 0, 1, 2, \ldots, n).$$

Par conséquent, l'égalité vectorielle (5) est équivalente au système d'équations

$$p_1 y_j^{(n-1)} + p_2 y_j^{(n-2)} + \ldots + p_n y_j^{(0)} = -y_j^{(n)} \quad (j = 1, 2, \ldots, n), \quad (6)$$

qui permet, en général, de déterminer les coefficients inconnus $p_1, p_2, \ldots, p_n$.

Puisque la formule (4) entraîne

$$y^{(k)} = A y^{(k-1)}$$

$(k = 1, 2, \ldots, n)$, les coordonnées $y_1^{(k)}, y_2^{(k)}, \ldots, y_n^{(k)}$ du vecteur $y^{(k)}$ se calculent successivement d'après les formules

$$\left.\begin{aligned} y_i^{(1)} &= \sum_{j=1}^{n} a_{ij} y_j^{(0)}, \\ y_i^{(2)} &= \sum_{j=1}^{n} a_{ij} y_j^{(1)}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_i^{(n)} &= \sum_{j=1}^{n} a_{ij} y_j^{(n-1)} \quad (i = 1, 2, \ldots, n). \end{aligned}\right\} \qquad (7)$$

Ainsi, la détermination des coefficients p_j du polynôme caractéristique (1) par la méthode de Krylov consiste à résoudre un système linéaire (6) aux coefficients calculés d'après les formules (7), les coordonnées du vecteur initial

$$y^{(0)} = \begin{bmatrix} y_1^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix}$$

étant arbitraires. Si le système (6) ne possède qu'une seule solution, ses composantes $p_1, p_2, \ldots, p_n$ sont les coefficients du polynôme caractéristique (1). Cette solution peut s'obtenir, par exemple, par

la méthode de Gauss (chapitre VIII, § 3). Si la solution du système (6) n'est pas unique, le problème se complique [1]. Dans ce cas il est recommandé de changer le vecteur initial.

Exemple. Trouver par la méthode de Krylov le polynôme caractéristique de la matrice (cf. § 3)

$$A=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}.$$

Solution. Choisissons le vecteur initial

$$y^{(0)}=\begin{bmatrix}1\\0\\0\\0\end{bmatrix}.$$

En utilisant les formules (7), définissons les coordonnées des vecteurs

$$y^{(k)}=A^k y^{(0)} \qquad (k=1,\ 2,\ 3,\ 4).$$

On a

$$y^{(1)}=Ay^{(0)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}=\begin{bmatrix}1\\2\\3\\4\end{bmatrix};$$

$$y^{(2)}=Ay^{(1)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}1\\2\\3\\4\end{bmatrix}=\begin{bmatrix}30\\22\\18\\20\end{bmatrix};$$

$$y^{(3)}=Ay^{(2)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}30\\22\\18\\20\end{bmatrix}=\begin{bmatrix}208\\178\\192\\242\end{bmatrix};$$

$$y^{(4)}=Ay^{(3)}=\begin{bmatrix}1&2&3&4\\2&1&2&3\\3&2&1&2\\4&3&2&1\end{bmatrix}\begin{bmatrix}208\\178\\192\\242\end{bmatrix}=\begin{bmatrix}2108\\1704\\1656\\1992\end{bmatrix}.$$

27—01072

Composons le système (6):

$$\begin{bmatrix} y_1^{(3)} & y_1^{(2)} & y_1^{(1)} & y_1^{(0)} \\ y_2^{(3)} & y_2^{(2)} & y_2^{(1)} & y_2^{(0)} \\ y_3^{(3)} & y_3^{(2)} & y_3^{(1)} & y_3^{(0)} \\ y_4^{(3)} & y_4^{(2)} & y_4^{(1)} & y_4^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = - \begin{bmatrix} y_1^{(4)} \\ y_2^{(4)} \\ y_3^{(4)} \\ y_4^{(4)} \end{bmatrix},$$

qui dans notre cas s'écrit

$$\begin{bmatrix} 208 & 30 & 1 & 1 \\ 178 & 22 & 2 & 0 \\ 192 & 18 & 3 & 0 \\ 242 & 20 & 4 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = - \begin{bmatrix} 2108 \\ 1704 \\ 1656 \\ 1992 \end{bmatrix}.$$

D'où

$$\left.\begin{aligned} 208p_1 + 30p_2 + p_3 + p_4 &= -2108, \\ 178p_1 + 22p_2 + 2p_3 &= -1704, \\ 192p_1 + 18p_2 + 3p_3 &= -1656, \\ 242p_1 + 20p_2 + 4p_3 &= -1992. \end{aligned}\right\}$$

En résolvant ce système on obtient:

$$p_1 = -4; \quad p_2 = -40; \quad p_3 = -56; \quad p_4 = -20.$$

Par conséquent,

$$\det (\lambda E - A) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20,$$

ce qui coïncide avec le résultat fourni par la méthode de Danilevski (§ 3).

§ 7. Calcul des vecteurs propres par la méthode de Krylov

La méthode de Krylov rend simple la recherche des vecteurs propres correspondants [1].

Pour simplifier, nous nous bornerons au cas où les racines λ_1, λ_2, . . ., λ_n du polynôme caractéristique

$$D(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

sont distinctes. Supposons que les coefficients du polynôme (1) et ses racines soient calculés. On demande de trouver les vecteurs propres $x^{(1)}$, $x^{(2)}$, . . ., $x^{(n)}$ associés respectivement aux valeurs propres λ_1, λ_2, . . ., λ_n.

Soient $y^{(0)}$, $y^{(1)} = Ay^{(0)}$, . . ., $y^{(n-1)} = A^{n-1}y^{(0)}$ les vecteurs utilisés dans la méthode de Krylov pour chercher les coefficients p_i ($i = 1, 2, \ldots, n$). En décomposant le vecteur $y^{(0)}$ suivant ses

vecteurs propres $x^{(i)}$ $(i = 1, 2, \ldots, n)$, on aura:

$$y^{(0)} = c_1x^{(1)} + c_2x^{(2)} + \ldots + c_nx^{(n)}, \tag{2}$$

c_i $(i = 1, 2, \ldots, n)$ étant certains coefficients numériques. D'où, compte tenu du fait que

$$\begin{aligned} &Ax^{(i)} = \lambda_i x^{(i)}, \\ &A^2x^{(i)} = \lambda_i^2 x^{(i)}, \qquad (i = 1, 2, \ldots, n), \\ &\ldots\ldots\ldots \end{aligned}$$

on aura:

$$\begin{aligned} y^{(1)} &= c_1\lambda_1 x^{(1)} + c_2\lambda_2 x^{(2)} + \ldots + c_n\lambda_n x^{(n)}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y^{(n-1)} &= c_1\lambda_1^{n-1}x^{(1)} + c_2\lambda_2^{n-1}x^{(2)} + \ldots + c_n\lambda_n^{n-1}x^{(n)}. \end{aligned} \tag{3}$$

Soit

$$\varphi_i(\lambda) = \lambda^{n-1} + q_{1i}\lambda^{n-2} + \ldots + q_{n-1,\,i} \tag{4}$$

$(i = 1, 2, \ldots, n)$ un système de polynômes arbitraire. En formant une combinaison linéaire de vecteurs $y^{(n-1)}, y^{(n-2)}, \ldots, y^{(0)}$ aux coefficients $1, q_{1-i}, \ldots, q_{n-1,\,i}$, on aboutit, en vertu des relations (2) et (3), à

$$\begin{aligned} y^{(n-1)} &+ q_{1i}y^{(n-2)} + \ldots + q_{n-1,i}\, y^{(0)} = \\ &= c_1\varphi_i(\lambda_1)\, x^{(1)} + c_2\varphi_i(\lambda_2)\, x^{(2)} + \ldots + c_n\varphi_i(\lambda_n)\, x^{(n)}. \end{aligned} \tag{5}$$

Si l'on adopte

$$\varphi_i(\lambda) = \frac{D(\lambda)}{\lambda - \lambda_i} \qquad (i = 1, 2, \ldots, n), \tag{6}$$

on a évidemment

$$\varphi_i(\lambda_j) = 0 \text{ pour } i \neq j$$

et

$$\varphi_i(\lambda_i) = D'(\lambda_i) \neq 0.$$

Dans ces conditions, la formule (5) s'écrit

$$\begin{aligned} c_i\varphi_i(\lambda_i)\, x^{(i)} = y^{(n-1)} &+ q_{1i}y^{(n-2)} + \\ &+ \ldots + q_{n-1,i}\, y^{(0)} \qquad (i = 1, 2, \ldots, n). \end{aligned} \tag{7}$$

Ainsi, si $c_i \neq 0$, la combinaison linéaire des vecteurs obtenue $y^{(n-1)}, y^{(n-2)}, \ldots, y^{(0)}$ donne le vecteur propre $x^{(i)}$ à un facteur numérique près. Les coefficients $q_{j,i}$ $(j = 1, 2, \ldots, n-1)$ peuvent être aisément déterminés d'après le *schéma de Horner*

$$\left.\begin{aligned} q_{0i} &= 1, \\ q_{ji} &= \lambda_i q_{j-1,i} + p_j. \end{aligned}\right\}$$

§ 8. Méthode de Leverrier

Cette méthode de développement d'un déterminant caractéristique a à sa base les formules de Newton [3] pour les sommes des puissances des racines d'une équation algébrique.

Soient

$$\det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \tag{1}$$

un polynôme caractéristique de la matrice $A = [a_{ij}]$ et $\lambda_1, \lambda_2, \ldots, \lambda_n$ l'ensemble complet de ses racines où toute racine est prise autant de fois que l'indique son ordre de multiplicité.

Posons

$$s_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k \qquad (k = 0, 1, 2, \ldots, n).$$

Alors pour $k \leqslant n$ on a les *formules de Newton* [3]

$$s_k + p_1 s_{k-1} + \ldots + p_{k-1} s_1 = -k p_k \quad (k = 1, 2, \ldots, n). \tag{2}$$

D'où

$$\left.\begin{aligned} p_1 &= -s_1, \\ p_2 &= -\frac{1}{2}(s_2 + p_1 s_1), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ p_n &= -\frac{1}{n}(s_n + p_1 s_{n-1} + \ldots + p_{n-1} s_1) \end{aligned}\right\} \tag{3}$$

Si les sommes $s_1, s_2, \ldots, s_n$ sont connues, les formules (3) permettent de trouver de proche en proche les coefficients $p_1, p_2, \ldots, p_n$ du polynôme caractéristique (1).

Les sommes $s_1, s_2, \ldots, s_n$ se calculent de la façon suivante: pour s_1 on a (cf. chapitre X, § 12):

$$s_1 = \lambda_1 + \lambda_2 + \ldots + \lambda_n = \operatorname{Sp} A,$$

c'est-à-dire

$$s_1 = \sum_{i=1}^{n} a_{ii}. \tag{4}$$

Ensuite, on sait (chapitre XI, § 1) que $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ sont les valeurs propres de la matrice A^k. Donc

$$s_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k = \operatorname{Sp} A^k,$$

c'est-à-dire si

$$A^k = [a_{ij}^{(k)}],$$

il vient

$$s_k = \sum_{i=1}^{n} a_{ii}^{(k)}. \tag{5}$$

Les puissances $A^k = A^{k-1}A$ s'obtiennent par multiplication directe.

Ainsi, le schéma de développement d'un déterminant caractéristique suivant la méthode de Leverrier est très simple, à savoir: on calcule d'abord les A^k $(k = 1, 2, \ldots, n)$ qui sont les puissances de la matrice A donnée, puis on trouve les s_k respectifs, sommes des éléments des diagonales principales des matrices A^k et, enfin, d'après les formules (3) on détermine les coefficients recherchés p_i $(i = 1, 2, \ldots, n)$.

La méthode de Leverrier est très délicate, car elle impose le calcul des puissances élevées de la matrice donnée. Son mérite est dû à un schéma de calcul peu compliqué et à l'absence de cas particuliers.

E x e m p l e. Développer par la méthode de Leverrier le déterminant caractéristique de la matrice (cf. § 3).

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix},$$

S o l u t i o n. Formons les puissances A^k $(k = 2, 3, 4)$ de la matrice A. On a:

$$A^2 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix};$$

$$A^3 = \begin{bmatrix} 30 & 22 & 18 & 20 \\ 20 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix};$$

$$A^4 = \begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 2108 & 1704 & 1656 & 1992 \\ 1704 & 1388 & 1368 & 1656 \\ 1656 & 1368 & 1388 & 1704 \\ 1992 & 1656 & 1704 & 2108 \end{bmatrix}.$$

Constatons qu'il n'a pas fallu calculer complètement A^4, il a suffi d'obtenir les éléments diagonaux principaux de cette matrice.

D'où

$$s_1 = \text{Sp } A = 1 + 1 + 1 + 1 = 4;$$
$$s_2 = \text{Sp } A^2 = 30 + 18 + 18 + 30 = 96;$$
$$s_3 = \text{Sp } A^3 = 208 + 148 + 148 + 208 = 712;$$
$$s_4 = \text{Sp } A^4 = 2108 + 1388 + 1388 + 2108 = 6992.$$

D'après les formules (3) on a donc:

$$p_1 = -s_1 = -4;$$
$$p_2 = -\frac{1}{2}(s_2 + p_1 s_1) = -\frac{1}{2}(96 - 4\cdot 4) = -40;$$
$$p_3 = -\frac{1}{3}(s_3 + p_1 s_2 + p_2 s_1) = -\frac{1}{3}(712 - 4\cdot 96 - 40\cdot 4) = -56;$$
$$p_4 = -\frac{1}{4}(s_4 + p_1 s_3 + p_2 s_2 + p_3 s_1) =$$
$$= -\frac{1}{4}(6992 - 4\cdot 712 - 40\cdot 96 - 56\cdot 4) = -20.$$

Ainsi on aboutit au résultat déjà connu (cf. § 3):

$$\begin{vmatrix} \lambda-1 & -2 & -3 & -4 \\ -2 & \lambda-1 & -2 & -3 \\ -3 & -2 & \lambda-1 & -2 \\ -4 & -3 & -2 & \lambda-1 \end{vmatrix} = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20.$$

§ 9. Notion de la méthode des coefficients indéterminés

Le déterminant caractéristique peut être développé également en cherchant un nombre suffisamment grand de ses valeurs numériques.

Soit

$$D(\lambda) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \qquad (1)$$

le déterminant caractéristique de la matrice A, c'est-à-dire

$$D(\lambda) = \det(\lambda E - A).$$

Si dans l'égalité (1) on pose successivement $\lambda = 0, 1, 2, \ldots, n-1$, on obtient pour les coefficients p_i ($i = 1, 2, \ldots, n$) un système d'équations linéaires

$$\left.\begin{aligned} p_n &= D(0), \\ 1^n + p_1\cdot 1^{n-1} + \ldots + p_n &= D(1), \\ 2^n + p_1\cdot 2^{n-1} + \ldots + p_n &= D(2), \\ &\ldots\ldots\ldots\ldots \\ (n-1)^n + p_1(n-1)^{n-1} + \ldots + p_n &= D(n-1). \end{aligned}\right\} \qquad (2)$$

D'où

$$\left.\begin{aligned} p_1+p_2+\ldots+p_{n-1} &= D(1)-D(0)-1, \\ 2^{n-1}p_1+2^{n-2}p_2+\ldots+2p_{n-1} &= D(2)-D(0)-2^n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ (n-1)^{n-1}p_1+(n-1)^{n-2}p_2+\ldots+(n-1)\ p_{n-1} &= \\ = D(n-1)-D(0)-(n-1) & \end{aligned}\right\} \qquad (3)$$

et

$$p_n = D(0) = \det(-A).$$

Le système (3) permet de calculer les coefficients p_i $(i = 1, 2, \ldots, n)$ du polynôme caractéristique (1).

Si l'on introduit la matrice

$$C_n = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 2^{n-1} & 2^{n-2} & \ldots & 2 \\ \ldots & \ldots & \ldots & \ldots \\ (n-1)^{n-1} & (n-1)^{n-2} & \ldots & n-1 \end{bmatrix}$$

et les vecteurs

$$D = \begin{bmatrix} D(1)-D(0)-1^n \\ D(2)-D(0)-2^n \\ \ldots\ldots\ldots\ldots \\ D(n-1)-D(0)-(n-1)^n \end{bmatrix}, \quad P = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \end{bmatrix}$$

le système (3) peut se mettre sous forme d'une équation matricielle

$$C_n P = D; \qquad (4)$$

d'où

$$P = C_n^{-1} D. \qquad (5)$$

Remarquons que l'inverse C_n^{-1} ne dépend que de l'ordre n du déterminant caractéristique et peut être établie à l'avance dans le cas d'un très grand nombre de déterminants caractéristiques de même ordre.

Ainsi l'application de cette méthode se ramène au calcul des déterminants numériques

$$D(k) = \det(kE - A) \quad (k = 0, 1, 2, \ldots, n-1)$$

et à la recherche de la solution d'un système linéaire standard (4).

§ 10. Comparaison de diverses méthodes de développement d'un déterminant caractéristique

Le tableau 26 [4] permet de juger de l'efficacité relative de diverses méthodes de développement du déterminant caractéristique. Ce tableau indique le nombre d'opérations qu'impose chacune des méthodes considérées en fonction de l'ordre du déterminant.

Tableau 26

Nombres d'opérations à effectuer dans le cas de diverses méthodes du développement du déterminant caractéristique en fonction de l'ordre du déterminant

Méthode	Ordre 3		4		5		7		9	
	Multiplications-divisions M-D	Additions-soustractions A-S	M-D	A-S	M-D	A-S	M-D	A-S	M-D	A-S
Développement direct	12	10	60	46	320	238	13 692	10 078	986 400	725 758
Danilevski	14	12	42	36	92	80	282	252	632	576
Krylov	67	38	179	118	389	280	1 287	1 022	3 209	2 688
Leverrier	41	27	153	114	414	330	1 791	1 533	5 228	4 644
Coefficients indéterminés	67	41	171	116	364	265	1 189	945	2 966	2 481
Interpolation *	46	38	125	102	279	230	972	826	2 525	2 202

* Cf. chapitre XIV, § 23.

On voit de ce tableau que pour développer les déterminants d'un ordre supérieur à cinq, c'est la méthode de Danilevski qui est la plus efficace du point de vue du nombre d'opérations.

§ 11. Calcul de la valeur propre la plus grande en module d'une matrice et d'un vecteur propre associé

Soit l'équation caractéristique

$$\det (A - \lambda E) = 0.$$

Les racines de cette équation $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont les valeurs propres de la matrice A. Supposons qu'à ces valeurs propres soient associés les vecteurs propres linéairement indépendants $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. Indiquons quelques méthodes itératives pour calculer la valeur

propre la plus grande en module d'une matrice A qui n'imposent pas le développement de son déterminant caractéristique.

C a s 1. Parmi les valeurs propres de la matrice A il y en a une seule la plus grande en module. Supposons, pour fixer les idées, que

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|. \quad (1)$$

Ainsi la plus grande en module est la p r e m i è r e v a l e u r p r o p r e. Evidemment, pour une matrice réelle la valeur propre λ_1 la plus grande en module est réelle. Notons que ce cas a lieu si la matrice A est réelle et si ses éléments sont positifs (chapitre X, § 16, théorème de Perron).

Indiquons un mode approché de calcul de la racine λ_1. Prenons un vecteur arbitraire y et développons-le par rapport aux vecteurs propres $x^{(j)}$ de la matrice A:

$$y = \sum_{j=1}^{n} c_j x^{(j)},$$

où c_j $(j = 1, 2, \ldots, n)$ sont des constantes. En appliquant la transformation A au vecteur y, on aura:

$$Ay = \sum_{j=1}^{n} c_j A x^{(j)}.$$

$x^{(j)}$ étant le vecteur propre de la transformation A, c'est-à-dire $Ax^{(j)} = \lambda_j x^{(j)}$, on en tire:

$$Ay = \sum_{j=1}^{n} c_j \lambda_j x^{(j)};$$

appelons Ay *itération* du vecteur y.

En composant successivement les itérations Ay, A^2y, ... A^my on tombe sur

$$A^m y = \sum_{j=1}^{n} c_j \lambda_j^m x^{(j)} \quad (2)$$

(*m-ième itération*).

Choisissons dans l'espace $E_n = \{y\}$ une base $e_1, e_2, \ldots, e_n$ quelconque. Soit

$$A^m y = y^{(m)} \qquad (m = 1, 2, 3, \ldots)$$

et

$$y^{(m)} = \begin{bmatrix} y_1^{(m)} \\ \vdots \\ y_n^{(m)} \end{bmatrix},$$

$y_i^{(m)}$ $(i = 1, 2, \ldots, n)$ étant les coordonnées du vecteur $y^{(m)}$ dans la base retenue.

Le développement des vecteurs propres $x^{(j)}$ par rapport aux vecteurs de la base donne

$$x^{(j)} = \sum_{j=1}^{n} x_{ij} e_i. \tag{3}$$

En portant (3) dans la formule (2), on a:

$$y^{(m)} = \sum_{j=1}^{n} c_j \lambda_j^m \sum_{i=1}^{n} x_{ij} e_i,$$

ou, en changeant l'ordre de sommation,

$$y^{(m)} = \sum_{i=1}^{n} e_i \sum_{j=1}^{n} c_j x_{ij} \lambda_j^m. \tag{4}$$

Le coefficient de e_i est la i-ème coordonnée du vecteur $y^{(m)}$. On peut donc écrire:

$$y_i^{(m)} = \sum_{j=1}^{n} c_j x_{ij} \lambda_j^m. \tag{4'}$$

D'une façon analogue

$$y_i^{(m+1)} = \sum_{j=1}^{n} c_j x_{ij} \lambda_j^{m+1}. \tag{4''}$$

La division de la deuxième somme par la première amène

$$\frac{y_i^{(m+1)}}{y_i^{(m)}} = \frac{c_1 x_{i1} \lambda_1^{m+1} + \ldots + c_n x_{in} \lambda_n^{m+1}}{c_1 x_{i1} \lambda_1^m + \ldots + c_n x_{in} \lambda_n^m}. \tag{5}$$

Supposons que $c_1 \neq 0$ et $x_{i1} \neq 0$. On peut l'obtenir en choisissant convenablement le vecteur initial y et la base $(e_1, e_2, \ldots, e_n)$.

Transformons l'expression (5) de la façon suivante:

$$\frac{y_i^{(m+1)}}{y_i^{(m)}} = \lambda_1 \frac{1 + \frac{c_2 x_{i2}}{c_1 x_{i1}} \left(\frac{\lambda_2}{\lambda_1}\right)^{m+1} + \ldots + \frac{c_n x_{in}}{c_1 x_{i1}} \left(\frac{\lambda_n}{\lambda_1}\right)^{m+1}}{1 + \frac{c_2 x_{i2}}{c_1 x_{i1}} \left(\frac{\lambda_2}{\lambda_1}\right)^{m} + \ldots + \frac{c_n x_{in}}{c_1 x_{i1}} \left(\frac{\lambda_n}{\lambda_1}\right)^{m}}.$$

En passant à la limite quand $m \to \infty$ et en tenant compte de l'inégalité (1), on a:

$$\lim_{m \to \infty} \frac{y_i^{(m+1)}}{y_i^{(m)}} = \lambda_1 \tag{6}$$

$\left(\text{puisque } \lim\limits_{m \to \infty} \left(\frac{\lambda_j}{\lambda_1}\right)^m = 0 \text{ pour } j > 1\right)$ ou, approximativement,

$$\lambda_1 \approx \frac{y_i^{(m+1)}}{y_i^{(m)}} \qquad (i = 1, 2, \ldots, n), \tag{7}$$

et, plus précisément,

$$\lambda_1 = \frac{y_i^{(m+1)}}{y_i^{(m)}} + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^m\right).$$

Si le numéro de l'itération est suffisamment grand, nous pouvons définir d'après la formule (7), avec une précision quelconque, la racine λ_1 la plus grande en module de l'équation caractéristique de la matrice A donnée. Pour chercher cette racine on peut utiliser une coordonnée quelconque du vecteur $y^{(m)}$; on peut prendre en particulier la moyenne arithmétique des rapports respectifs.

R e m a r q u e 1. Dans les cas exceptionnels, lorsque le choix du vecteur y est mauvais, il se peut que la formule (6) ne donne pas la racine cherchée ou même n'ait aucun sens, c'est-à-dire il se peut que la limite du rapport $\frac{y_i^{(m+1)}}{y_i^{(m)}}$ n'existe pas. On s'en aperçoit facilement d'après les valeurs « sautantes » de ce rapport. Il faut alors essayer un autre vecteur initial.

R e m a r q u e 2. Pour accélérer la convergence de l'itération (6), il est quelquefois avantageux de composer la suite des matrices

$$A^2 = A \cdot A,$$
$$A^4 = A^2 \cdot A^2,$$
$$A^8 = A^4 \cdot A^4,$$
$$\cdots\cdots\cdots$$
$$A^{2^k} = A^{2^{k-1}} \cdot A^{2^{k-1}}.$$

D'où l'on tire

$$y^{(m)} = A^m y$$

et

$$y^{(m+1)} = A y^{(m)},$$

avec $m = 2^k$. Ensuite, on pose comme d'habitude:

$$\lambda_1 \approx \frac{y_i^{(m+1)}}{y_i^{(m)}} \qquad (i = 1, 2, \ldots, n).$$

Le vecteur $y^{(m)} = A^m y$ est approximativement le vecteur propre de la matrice A, associé à la valeur propre λ_1. En effet, la formule (2) entraîne:

$$A^m y = c_1 \lambda_1^m x^{(1)} + \sum_{j=2}^{n} c_j \lambda_j^m x^{(j)},$$

où $x^{(j)}$ $(j = 1, 2, \ldots, n)$ sont les vecteurs propres de A.

Par suite

$$A^m y = c_1\lambda_1^m \left\{x^{(1)} + \sum_{j=2}^{n} \frac{c_j}{c_1}\left(\frac{\lambda_j}{\lambda_1}\right)^m x^{(j)}\right\}.$$

Comme $\left(\frac{\lambda_j}{\lambda_1}\right)^m \to 0$ quand $m \to \infty$ $(j > 1)$, pour m suffisamment grand on aura avec la précision qu'on voudra

$$A^m y \approx c_1\lambda_1^m x^{(1)},$$

c'est-à-dire que $A^m y$ ne se distingue du vecteur propre $x^{(1)}$ que par le facteur numérique et, par conséquent, il est également un vecteur propre associé à la même valeur propre λ_1.

E x e m p l e. Trouver la plus grande valeur propre de la matrice

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \qquad (8)$$

et le vecteur propre qui lui correspond.

S o l u t i o n. Choisissons un vecteur initial

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Composons le tableau 27.

Tableau 27

Calcul de la première valeur propre

y	Ay	A^2y	A^3y	A^4y	A^5y	A^6y	A^7y	A^8y	A^9y	$A^{10}y$
1	5	24	111	504	2268	10 161	45 433	202 833	905 238	4 038 939
1	4	15	60	252	1089	4 779	21 141	93 906	417 987	1 862 460
1	2	6	21	81	333	1 422	6 201	27 342	121 248	539 235

En s'arrêtant aux itérations $A^9 y = y^{(9)}$ et $A^{10}y = y^{(10)}$, on obtient les valeurs

$$\frac{y_1^{(10)}}{y_1^{(9)}} = \frac{4038939}{905238} = 4{,}462;$$

$$\frac{y_2^{(10)}}{y_2^{(9)}} = \frac{1862460}{417987} = 4{,}456;$$

$$\frac{y_3^{(10)}}{y_3^{(9)}} = \frac{539235}{121248} = 4{,}447.$$

On peut donc poser approximativement :

$$\lambda_1 = \frac{1}{3}(4{,}462 + 4{,}456 + 4{,}447) = 4{,}455 \approx 4{,}46.$$

Comme premier vecteur propre de A on peut prendre

$$A^{10}y = \begin{bmatrix} 4038939 \\ 1862460 \\ 539235 \end{bmatrix}.$$

Après sa normalisation on obtient finalement :

$$x^{(1)} = \begin{bmatrix} 0{,}90 \\ 0{,}42 \\ 0{,}12 \end{bmatrix}.$$

Cas 2. La valeur propre de A la plus grande en module est multiple.

Soit

$$\lambda_1 = \lambda_2 = \ldots = \lambda_s$$

et

$$|\lambda_1| > |\lambda_k| \quad \text{pour} \quad k > s.$$

La formule (5) conduit à

$$\frac{y_i^{(m+1)}}{y_i^{(m)}} =$$

$$= \frac{c_1 x_{i1}\lambda_1^{m+1} + \ldots + c_s x_{is}\lambda_1^{m+1} + c_{s+1}x_{i,s+1}\lambda_{s+1}^{m+1} + \ldots + c_n x_{in}\lambda_n^{m+1}}{c_1 x_{i1}\lambda_1^{m} + \ldots + c_s x_{is}\lambda_1^{m} + c_{s+1}x_{i,s+1}\lambda_{s+1}^{m} + \ldots + c_n x_{in}\lambda_n^{m}} =$$

$$= \lambda_1 \frac{c_1 x_{i1} + \ldots + c_s x_{is} + c_{s+1}x_{i,s+1}\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^{m+1} + \ldots + c_n x_{in}\left(\frac{\lambda_n}{\lambda_1}\right)^{m+1}}{c_1 x_{i1} + \ldots + c_s x_{is} + c_{s+1}x_{i,s+1}\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^{m} + \ldots + c_n x_{in}\left(\frac{\lambda_n}{\lambda_1}\right)^{m}}.$$

D'où si $c_1 x_{i1} + \ldots + c_s x_{is} \neq 0$ et en tenant compte que

$$\left(\frac{\lambda_k}{\lambda_1}\right)^m \to 0 \text{ quand } m \to \infty \text{ et } k > s,$$

on obtient

$$\lim_{m\to\infty} \frac{y_i^{(m+1)}}{y_i^{(m)}} = \lambda_1 \qquad (i = 1, 2, \ldots, n)$$

ou plus précisément

$$\lambda_1 = \frac{y_i^{(m+1)}}{y_i^{(m)}} + O\left(\left(\frac{\lambda_{s+1}}{\lambda_1}\right)^m\right).$$

Donc dans ce cas-là aussi on peut appliquer le procédé de calcul de λ_1 décrit dans ce qui précède.

De même qu'auparavant,

$$y^{(m)} = A^m y$$

est l'un des vecteurs propres approchés de la matrice A associés à la valeur λ_1. En changeant le vecteur initial y nous obtenons, dans le cas général, un autre vecteur de A linéairement indépendant. Constatons que dans ce cas, pour la valeur λ_1, le procédé appliqué ne garantit pas la détermination de l'ensemble tout entier des vecteurs propres linéairement indépendants de la matrice A.

Pour les cas 1-2 on peut indiquer une procédure itérative plus rapide de la recherche de la valeur propre λ_1 la plus grande en module; composons la suite des matrices

$$A,\ A^2,\ A^4,\ A^8,\ \ldots,\ A^{2k}.$$

On sait (chapitre X, § 12) que

$$\sum_{i=1}^{n} \lambda_i = \operatorname{Sp} A;$$

d'une façon analogue

$$\sum_{i=1}^{n} \lambda_i^m = \operatorname{Sp} A^m,$$

avec $m = 2^k$. En nous bornant pour simplifier au cas 1, on a:

$$\lambda_1^m + \lambda_2^m + \ldots + \lambda_n^m = \lambda_1^m \left[1 + \left(\frac{\lambda_2}{\lambda_1}\right)^m + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^m\right] = \operatorname{Sp} A^m;$$

d'où

$$|\lambda_1| \left[1 + \left(\frac{\lambda_2}{\lambda_1}\right)^m + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^m\right]^{\frac{1}{m}} = \sqrt[m]{\operatorname{Sp} A^m}.$$

Quand $m \to \infty$, on obtient:

$$|\lambda_1| = \lim_{m \to \infty} \sqrt[m]{\operatorname{Sp} A^m},$$

c'est-à-dire

$$|\lambda_1| \approx \sqrt[m]{\operatorname{Sp} A^m},$$

où m est suffisamment grand.

Pour éviter l'extraction des racines de grands indices, on peut trouver

$$A^{m+1} = A^m A.$$

Il vient

$$\lambda_1^{m+1} + \lambda_2^{m+1} + \ldots + \lambda_n^{m+1} = \operatorname{Sp} A^{m+1}$$

et

$$\lambda_1^m+\lambda_2^m+\ldots+\lambda_n^m=\operatorname{Sp} A^m.$$

Il en résulte, compte tenu de la petitesse relative de $|\lambda_2|$, ..., $|\lambda_n|$ par rapport à $|\lambda_1|$,

$$\lambda_1\approx \operatorname{Sp} A^{m+1}/\operatorname{Sp} A^m.$$

§ 12. Application de la méthode des produits scalaires au calcul de la première valeur propre d'une matrice réelle

Le calcul de la première valeur propre λ_1 d'une matrice réelle A peut se faire en appliquant un autre processus itératif quelquefois plus avantageux. Cette méthode est basée sur la formation des produits scalaires

$$(A^k y_0,\ A'^k y_0)\quad \text{et}\quad (A^{k-1}y_0,\ A'^k y_0)$$

($k = 1, 2, \ldots$) où A' est une transposée de la matrice A et y_0 un vecteur initial choisi d'une façon quelconque.

Passons maintenant à l'exposé de cette méthode.

Soit A une matrice réelle et $\lambda_1, \lambda_2, \ldots, \lambda_n$ ses valeurs propres qu'on suppose distinctes et telles que

$$|\lambda_1|>|\lambda_2|\geqslant\ldots\geqslant|\lambda_n|.$$

Prenons un certain vecteur y_0 non nul et construisons à l'aide de la matrice A la suite des itérations

$$y_k = Ay_{k-1}\quad (k=1,\ 2,\ \ldots). \tag{1}$$

Formons également pour le vecteur y_0 à l'aide de la transposée A' une deuxième suite des itérations

$$y'_k=A'y'_{k-1}\quad (k=1,\ 2,\ \ldots,), \tag{2}$$

où $y'_0 = y_0$.

D'après le théorème 1 du chapitre X, § 16, choisissons dans l'espace E_n deux bases propres $\{x_j\}$ et $\{x'_j\}$ respectivement pour les matrices A et A' qui vérifient les conditions de biorthonormalisation:

$$(x_i,\ x'_j)=\delta_{ij}, \tag{3}$$

avec $Ax_i=\lambda_i x_i$ et $A'x'_j=\lambda_j^* x'_j$ ($i,\ j=1,\ 2,\ \ldots,n$). Désignons les coordonnées du vecteur y_0 dans la base $\{x_j\}$ par $a_1, a_2, \ldots, a_n$, et dans la base $\{x'_j\}$ par $b_1, b_2, \ldots, b_n$, c'est-à-dire

$$y_0=a_1x_1+\ldots+a_nx_n\quad \text{et}\quad y_0=b_1x'_1+\ldots+b_nx'_n.$$

D'où

$$y_k=A^k y_0=\sum_{j=1}^{n} a_j\lambda_j^k x_j \tag{4}$$

et

$$y'_k = A'^k y_0 = \sum_{j=1}^{n} b_j \lambda_j^{*k} x'_j \quad (k = 1, 2, \ldots). \tag{4'}$$

Composons le produit scalaire

$$(y_k, y'_k) = (A^k y_0, A'^k y_0) = (y_0, A'^{2k} y_0) = \\ = \left(\sum_{i=1}^{n} a_i x_i, \sum_{j=1}^{n} b_j \lambda_j^{*2k} x'_j \right).$$

La condition d'orthonormalisation entraîne

$$(y_k, y'_k) = \sum_{j=1}^{n} a_j b_j^* \lambda_j^{2k} = \\ = a_1 b_1^* \lambda_1^{2k} + a_2 b_2^* \lambda_2^{2k} + \ldots + a_n b_n^* \lambda_n^{2k}. \tag{5}$$

D'une façon analogue

$$(y_{k-1}, y'_k) = a_1 b_1^* \lambda_1^{2k-1} + a_2 b_2^* \lambda_2^{2k-1} + \ldots + a_n b_n^* \lambda_n^{2k-1}. \tag{6}$$

Par conséquent, pour $a_1 b_1^* \neq 0$, on a :

$$\frac{(y_k, y'_k)}{(y_{k-1}, y'_k)} = \frac{a_1 b_1^* \lambda_1^{2k} + a_2 b_2^* \lambda_2^{2k} + \ldots + a_n b_n^* \lambda_n^{2k}}{a_1 b_1^* \lambda_1^{2k-1} + a_2 b_2^* \lambda_2^{2k-1} + \ldots + a_n b_n^* \lambda_n^{2k-1}} = \lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1} \right)^{2k} \right).$$

Ainsi

$$\lambda_1 \approx \frac{(y_k, y'_k)}{(y_{k-1}, y'_k)} = \frac{(A^k y_0, A'^k y_0)}{(A^{k-1} y_0, A'^k y_0)} \tag{7}$$

Cette méthode est commode surtout pour une matrice symétrique A du fait qu'alors $A' = A$ et on a simplement

$$\lambda_1 \approx \frac{(A^k y_0, A^k y_0)}{(A^{k-1} y_0, A^k y_0)}; \tag{8}$$

il ne faut donc former qu'une seule suite $y_k = A^k y_0$ $(k = 1, 2, \ldots)$.

E x e m p l e. Chercher par la méthode des produits scalaires la plus grande valeur propre de la matrice (§ 11)

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

S o l u t i o n. La matrice A étant symétrique, il suffit de construire une seule suite d'itérations $A^k y_0$ $(k = 1, 2, \ldots)$. En adoptant pour vecteur initial

$$y_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

on peut utiliser les résultats fournis par le tableau 27. Par exemple, pour $k = 5$ et $k = 6$, on a

$$A^5 y_0 = \begin{bmatrix} 2\ 268 \\ 1\ 089 \\ 333 \end{bmatrix} \quad \text{et} \quad A^6 y_0 = \begin{bmatrix} 10\ 161 \\ 4\ 779 \\ 1\ 422 \end{bmatrix}.$$

D'où

$$(A^5 y_0, A^6 y_0) = 2268 \cdot 10161 + 1089 \cdot 4779 + 333 \cdot 1422 = 28\ 723\ 005$$

et

$$(A^6 y_0, A^6 y_0) = 10\ 161^2 + 4779^2 + 1422^2 = 128\ 106\ 846.$$

Par suite,

$$\lambda_1 \approx \frac{(A^6 y_0, A^6 y_0)}{(A^5 y_0, A^6 y_0)} = \frac{128\ 106\ 846}{28\ 723\ 005} = 4{,}46,$$

ce qui coïncide, pour les chiffres écrits, avec la valeur obtenue au § 11 à l'aide de $A^{10} y_0$.

R e m a r q u e. Les méthodes de calcul de la racine la plus grande en module d'une équation caractéristique (§ 11) peuvent être utilisées pour le calcul de la racine la plus grande en module d'une équation algébrique

$$x^n + p_1 x^{n-1} + \ldots + p_n = 0. \tag{9}$$

En effet, on vérifie immédiatement que l'équation (9) est caractéristique pour la matrice (cf. § 3, matrice de Frobénius)

$$P = \begin{bmatrix} -p_1 & -p_2 \ldots & -p_{n-1} & -p_n \\ 1 & 0 \ \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 \ \ldots & 1 & 0 \end{bmatrix},$$

c'est-à-dire (9) est équivalente à l'équation

$$\det (xP - E) = 0.$$

Si l'équation (9) ne possède pas de racines nulles, on peut déterminer d'une façon analogue la racine la plus petite en module de cette équation, et notamment pour $p_n \neq 0$, en posant $\frac{1}{x} = y$, on obtient

$$y^n + \frac{p_{n-1}}{p_n} y^n + \ldots + \frac{1}{p_n} = 0. \tag{10}$$

La valeur inverse de la racine la plus grande en module de (10) donne évidemment la racine la plus petite en module de (9).

§ 13. Calcul de la deuxième valeur propre et du deuxième vecteur propre d'une matrice

Supposons que les valeurs propres de λ_j $(j = 1, 2, \ldots, n)$ de la matrice A sont telles que

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|, \tag{1}$$

c'est-à-dire qu'il existe deux valeurs propres distinctes λ_1 et λ_2 de la matrice A les plus grandes en module. Dans ce cas en appliquant le procédé analogue à celui du § 11 on peut calculer approximativement la deuxième valeur propre λ_2 et le vecteur propre $x^{(2)}$ associé.

La formule (2) du § 11 entraîne

$$A^m y = c_1\lambda_1^m x^{(1)} + c_2\lambda_2^m x^{(2)} + \ldots + c_n\lambda_n^m x^{(n)} \tag{2}$$

et

$$A^{m+1} y = c_1\lambda_1^{m+1} x^{(1)} + c_2\lambda_2^{m+1} x^{(2)} + \ldots + c_n\lambda_n^{m+1} x^{(n)}. \tag{3}$$

Eliminons des formules (2) et (3) les termes contenant λ_1. A cette fin retranchons de l'égalité (3) le produit de l'égalité (2) par λ_1. Il en résulte

$$A^{m+1} y - \lambda_1 A^m y = c_2\lambda_2^m (\lambda_2 - \lambda_1) x^{(2)} + \ldots + c_n\lambda_n^m (\lambda_n - \lambda_1) x^{(n)}. \tag{4}$$

Pour abréger l'écriture introduisons les notations

$$\Delta_\lambda A^m y = A^{m+1} y - \lambda A^m y; \tag{5}$$

appelons l'expression (5) *λ-différence* de $A^m y$. Si $c_2 \neq 0$, il est évident que le premier terme du deuxième membre de (4) est son terme principal quand $m \to \infty$, et nous avons l'égalité approchée

$$\Delta_{\lambda_1} A^m y \approx c_2\lambda_2^m (\lambda_2 - \lambda_1) x^{(2)}. \tag{6}$$

D'où

$$\Delta_{\lambda_1} A^{m-1} y \approx c_2\lambda_2^{m-1} (\lambda_2 - \lambda_1) x^{(2)}. \tag{7}$$

Soit

$$A^m y = y^{(m)} = \begin{bmatrix} y_1^{(m)} \\ y_2^{(m)} \\ \vdots \\ y_n^{(m)} \end{bmatrix}.$$

Les formules (6) et (7) donnent

$$\lambda_2 \approx \frac{\Delta_{\lambda_1} y_i^{(m)}}{\Delta_{\lambda_1} y_i^{(m-1)}} = \frac{y_i^{(m+1)} - \lambda_1 y_i^{(m)}}{y_i^{(m)} - \lambda_1 y_i^{(m-1)}} \quad (i = 1, 2, \ldots, n). \tag{8}$$

L'utilisation de la formule (8) permet d'obtenir par calcul approché la deuxième valeur propre λ_2. Remarquons qu'en pratique, vu

les pertes de précision par soustraction de nombres voisins, il est quelquefois plus avantageux de prendre le numéro de l'itération k pour λ_2 plus petit que le numéro de l'itération m pour λ_1, c'est-à-dire il est rationnel d'adopter:

$$\lambda_2 \approx \frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}} \quad (k < m), \tag{9}$$

où k est le plus petit nombre pour lequel la domination de λ_2 sur les valeurs propres successives devient manifeste. En général, la formule (9) donne des valeurs grossières de λ_2. Constatons que si les modules de toutes les valeurs propres sont distincts, les formules analogues à (9) permettent de calculer également les autres valeurs propres de la matrice donnée. Mais les résultats de ces calculs seront encore moins sûrs.

Pour ce qui est du vecteur propre $x^{(2)}$, la formule (6) montre qu'on peut poser

$$x^{(2)} \approx \Delta_{\lambda_1} y^{(k)}. \tag{10}$$

Il existe une extension de cette méthode au cas des racines multiples d'une équation caractéristique [1].

Exemple. Déterminer les valeurs propres et les vecteurs propres successifs de la matrice (cf. exemple de § 11)

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Solution. Pour calculer la deuxième valeur propre adoptons $k = 8$. On a (cf. tableau 27):

$A^7 y$	$A^8 y$	$A^9 y$
45 433	202 833	905 238
21 141	93 906	417 987
6 201	27 342	121 248

Composons les λ-différences d'après la formule

$$\Delta_{\lambda_1} y_i^{(j)} = y_i^{(j+1)} - \lambda_1 y_i^{(j)} \quad (i = 1, 2, 3),$$

avec $y^{(j)} = A^j y$. Pour chacune des colonnes on adopte sa valeur de λ_1, soit $\lambda_1 = 4{,}462$; $\lambda_1 = 4{,}456$; $\lambda_1 = 4{,}447$ (tableau 28).

Tableau 28

Calcul de la deuxième valeur propre

A^8y	$\lambda_1 A^7 y$	$\Delta_{\lambda_1} A^7 y$	$A^9 y$	$\lambda_1 A^8 y$	$\Delta_{\lambda_1} A^8 y$
202 833 93 906 27 342	202 722 94 204 27 576	111 −298 −234	905 238 417 987 121 248	905 041 418 445 121 590	197 −458 −342

On en tire

$$\frac{\Delta_{\lambda_1} y_1^{(8)}}{\Delta_{\lambda_1} y_1^{(7)}} = \frac{197}{111} = 1,78; \quad \frac{\Delta_{\lambda_1} y_2^{(8)}}{\Delta_{\lambda_1} y_2^{(7)}} = \frac{-458}{-298} = 1,54;$$

$$\frac{\Delta_{\lambda_1} y_3^{(8)}}{\Delta_{\lambda_1} y_3^{(7)}} = \frac{-342}{-234} = 1,46.$$

On peut donc poser approximativement :

$$\lambda_2 = \frac{1}{3}(1,78 + 1,54 + 1,46) \approx 1,59.$$

Pour deuxième vecteur propre on adopte :

$$\Delta_{\lambda_1} A^8 y = \begin{bmatrix} 197 \\ -458 \\ -342 \end{bmatrix}.$$

La normalisation de ce vecteur donne :

$$x^{(2)} = \begin{bmatrix} 0,33 \\ -0,76 \\ -0,56 \end{bmatrix}.$$

La matrice A étant symétrique, les vecteurs $x^{(1)}$ (§ 11) et $x^{(2)}$ doivent être orthogonaux entre eux. La vérification donne :

$$(x^{(1)}, x^{(2)}) = 0,90 \cdot 0,33 + 0,42 \cdot (-0,76) + 0,12 \cdot (-0,56) = 0,09.$$

D où $\measuredangle (x^{(1)}, x^{(2)}) = 85°$, ce qui est assez imprécis.

La troisième valeur propre λ_3 se trouve d'après la trace de A :

$$\lambda_1 + \lambda_2 + \lambda_3 = \mathrm{Sp}\, A = 4 + 2 + 1 = 7.$$

Il en résulte

$$\lambda_3 = 7 - 4,46 - 1,59 \approx 0,95.$$

Le vecteur propre

$$x^{(3)} = \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix}$$

se calcule à partir des conditions d'orthogonalité

$$\left.\begin{aligned} 0{,}90x_1^{(3)} + 0{,}42x_2^{(3)} + 0{,}12x_3^{(3)} &= 0, \\ 0{,}33x_1^{(3)} + (-0{,}76)\,x_2^{(3)} + (-0{,}56)\,x_3^{(3)} &= 0. \end{aligned}\right\}$$

Il en résulte

$$\frac{x_1^{(3)}}{\begin{vmatrix} 0{,}42 & 0{,}12 \\ -0{,}76 & -0{,}56 \end{vmatrix}} = \frac{x_2^{(3)}}{\begin{vmatrix} 0{,}12 & 0{,}90 \\ -0{,}56 & 0{,}33 \end{vmatrix}} = \frac{x_3^{(3)}}{\begin{vmatrix} 0{,}90 & 0{,}42 \\ 0{,}33 & -0{,}76 \end{vmatrix}}$$

ou

$$\frac{x_1^{(3)}}{-0{,}144} = \frac{x_2^{(3)}}{0{,}539} = \frac{x_3^{(3)}}{-0{,}818}\,.$$

Après normalisation on obtient finalement :

$$x^{(3)} = \begin{bmatrix} -0{,}14 \\ 0{,}53 \\ -0{,}81 \end{bmatrix}.$$

§ 14. Méthode d'exhaustion

Il existe encore une méthode pour déterminer la deuxième valeur propre d'une matrice et son vecteur propre associé, dite ***méthode d'exhaustion*** [1].

Supposons que la matrice $A = [a_{ij}]$ soit réelle et possède des valeurs propres distinctes $\lambda_1, \lambda_2, \ldots, \lambda_n$, de plus

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|.$$

Considérons avec la matrice A la matrice

$$A_1 = A - \lambda_1 X_1 X_1', \tag{1}$$

où λ_1 est la première valeur propre de A,

$$X_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix}$$

est le vecteur propre correspondant de A considéré comme matrice colonne, et

$$X_1' = [x_{11}' x_{21}' \ldots x_{n1}']$$

est le vecteur propre associé à λ_1 de la transposée A', considérée comme matrice ligne; en outre, les vecteurs X_1 et X_1' sont normalisés de façon que leur produit scalaire soit égal à un:

$$(X_1, X_1'^*) = X_1'X_1 = \sum_{j=1}^{n} x_{j1}x'_{j1} = 1. \tag{2}$$

Le nombre λ_1 et les vecteurs X_1 et X_1' sont supposés connus.

La matrice A_1 s'écrit sous forme développée

$$A_1 = \begin{bmatrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \dots & \dots \\ a_{n1} & a_{n2} \dots a_{nn} \end{bmatrix} - \lambda_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} [x'_{11}x'_{21} \dots x'_{n1}] =$$

$$= \begin{bmatrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \dots & \dots \\ a_{n1} & a_{n2} \dots a_{nn} \end{bmatrix} - \lambda_1 \begin{bmatrix} x_{11}x'_{11} & x_{11}x'_{21} \dots x_{11}x'_{n1} \\ x_{21}x'_{11} & x_{21}x'_{21} \dots x_{21}x'_{n1} \\ \dots & \dots \\ x_{n1}x'_{11} & x_{n1}x'_{21} \dots x_{n1}x'_{n1} \end{bmatrix}. \tag{1'}$$

Montrons que tous vecteurs propres X_j ($j = 1, 2, \dots, n$) de A sont aussi vecteurs propres de A_1; de plus, les valeurs propres associées sont conservées, sauf λ_1 qui est remplacée par une valeur propre nulle.

En effet, l'associativité d'un produit matriciel et la condition de normalisation (2) font qu'on a

$$A_1X_1 = AX_1 - \lambda_1 (X_1X_1') X_1 = \lambda_1X_1 - \\ - \lambda_1X_1 (X_1'X_1) = \lambda_1X_1 - \lambda_1X_1 = 0,$$

c'est-à-dire

$$A_1X_1 = 0X_1$$

et donc la valeur propre de A est zéro.

Ensuite, pour $j > 1$ et tenant compte du fait que

$$(X_j, X_1'^*) = X_1'X_j = 0 \quad (j = 2, \dots, n)$$

(cf. chapitre X, § 16, théorème 1), on obtient

$$A_1X_j = AX_j - \lambda_1 (X_1X_1') X_j = \lambda_1X_j - \lambda_1X_1 (X_1'X_j) = \lambda_jX_j \\ (j = 2, \dots, n).$$

Ainsi, pour la matrice A_1 la valeur propre la plus grande en module est λ_2. Pour déterminer λ_2 et le vecteur propre X_2 associé on peut donc faire appel aux méthodes indiquées aux §§ 11 et 12. Ce procédé s'appelle *méthode d'exhaustion*. Par exemple, en partant

du vecteur arbitraire y_0 on peut calculer λ_2 d'après la formule

$$\lambda_2 \approx \frac{(A_1^m y_0)_i}{(A_1^{m-1} y_0)_i} \qquad (i=1, 2, \ldots, n),$$

de plus

$$X_2 \approx cA_1^m y_0 \quad (c \neq 0).$$

Montrons que pour chercher les itérations $A_1^m y_0$ ($m = 1, 2, \ldots$) on peut utiliser la formule

$$A_1^m y_0 = A^m y_0 - \lambda_1^m X_1 X_1' y_0, \tag{3}$$

qui permet d'éviter l'itération directe de la matrice A_1.

En effet, supposons que les vecteurs propres X_j et X_j' ($j = 1, 2, \ldots, n$) de A et de la transposée A' vérifient les conditions de biorthonormalisation (chapitre X, § 16, théorème 2)

$$X_k' X_j = \delta_{jk},$$

où δ_{jk} est le symbole de Kronecker. On est alors en présence d'un développement bilinéaire de A:

$$A = \lambda_1 X_1 X_1' + \lambda_2 X_2 X_2' + \ldots + \lambda_n X_n X_n'. \tag{4}$$

D'où

$$A_1 = A - \lambda_1 X_1 X_1' = \lambda_2 X_2 X_2' + \ldots + \lambda_n X_n X_n'. \tag{5}$$

Comme

$$A^m X_j = \lambda_j^m X_j \quad (j=1, 2, \ldots, n),$$

en prémultipliant (4) par A^{m-1}, on a:

$$A^m = A^m X_1 X_1' + A^m X_2 X_2' + \ldots + A^m X_n X_n' =$$
$$= \lambda_1^m X_1 X_1' + \lambda_2^m X_2 X_2' + \ldots + \lambda_n^m X_n X_n'. \tag{6}$$

D'une façon analogue, en tenant compte de

$$A_1^m X_1 = A_1^{m-1} (A_1 X_1) = 0$$

et

$$A_1^m X_j = \lambda_j^m X_j \quad (j=2, 3, \ldots, n),$$

on obtient après prémultiplication de (5) par A_1^{m-1}

$$A_1^m = A_1^m X_2 X_2' + \ldots + A_1^m X_n X_n' =$$
$$= \lambda_2^m X_2 X_2' + \ldots + \lambda_n^m X_n X_n'. \tag{7}$$

Les formules (6) et (7) entraînent

$$A_1^m = A^m - \lambda_1^m X_1 X_1',$$

ce qui est équivalent à la relation (3).

§ 15. Calcul des éléments propres d'une matrice symétrique définie positive

Exposons la méthode itérative de la recherche simultanée des valeurs propres et des vecteurs propres d'une matrice définie positive [5].

On sait (chapitre X, § 15) que si une matrice réelle

$$A = [a_{ij}]$$

est symétrique et définie positive,

1) les racines $\lambda_1, \lambda_2, \ldots, \lambda_n$ de son équation caractéristique

$$\begin{bmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \ldots & a_{nn}-\lambda \end{bmatrix} = 0 \qquad (1)$$

sont réelles et positives;

2) les vecteurs propres

$$x^{(j)} = \begin{bmatrix} x_1^{(j)} \\ \vdots \\ x_n^{(j)} \end{bmatrix} \qquad (j = 1, 2, \ldots n)$$

peuvent être pris réels et vérifient les conditions d'orthogonalité

$$\sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} = 0 \quad \text{pour } j \neq k. \qquad (2)$$

Ecrivons un système qui permet de calculer le vecteur propre $x^{(1)}$:

$$\left.\begin{aligned} (a_{11}-\lambda_1)\, x_1^{(1)} + a_{12}x_2^{(1)} + \ldots + a_{1n}x_n^{(1)} &= 0, \\ a_{21}x_1^{(1)} + (a_{22}-\lambda_1)\, x_2^{(1)} + \ldots + a_{2n}x_n^{(1)} &= 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \ldots + (a_{nn}-\lambda_1)\, x_n^{(1)} &= 0 \end{aligned}\right\}$$

ou

$$\left.\begin{aligned} x_1^{(1)} &= \frac{1}{\lambda_1}\left(a_{11}x_1^{(1)} + a_{12}x_2^{(1)} + \ldots + a_{1n}x_n^{(1)}\right), \\ x_2^{(1)} &= \frac{1}{\lambda_1}\left(a_{21}x_1^{(1)} + a_{22}x_2^{(1)} + \ldots + a_{2n}x_n^{(1)}\right), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{n-1}^{(1)} &= \frac{1}{\lambda_1}\left(a_{n-1,1}x_1^{(1)} + a_{n-1,2}x_2^{(1)} + \ldots + a_{n-1,n}x_n^{(1)}\right), \\ \lambda_1 &= \frac{1}{x_n^{(1)}}\left(a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + \ldots + a_{nn}x_n^{(1)}\right). \end{aligned}\right\} \qquad (3)$$

Les coordonnées des vecteurs propres étant définies à un facteur de proportionnalité près, l'une d'elles est arbitraire; par exemple, on peut poser, sauf le cas particulier, $x_n^{(1)} = 1$. En général, le système (3) peut être résolu par la méthode itérative [5] en choisissant des valeurs initiales convenables $x_i^{(1,0)}$, $\lambda_1^{(0)}$ et en posant

$$x_i^{(1,k+1)} = \frac{1}{\lambda_1^{(k)}} \left(\sum_{j=1}^{n-1} a_{ij} x_j^{(1,k)} + a_{in} \right) \quad (i = 1, 2, \ldots, n-1);$$

$$\lambda_1^{(k+1)} = \sum_{j=1}^{n-1} a_{nj} x_j^{(1,k+1)} + a_{nn} \quad (k = 0, 1, 2, \ldots).$$

On peut également utiliser le processus de Seidel. C'est ainsi qu'on obtient la première racine de (1)

$$\lambda_1 \approx \lambda_1^{(k)} \tag{4}$$

et le premier vecteur propre

$$x^{(1)} \approx \begin{bmatrix} x_1^{(1,k)} \\ \vdots \\ x_{n-1}^{(1,k)} \\ 1 \end{bmatrix}.$$

Pour calculer la deuxième racine λ_2 de (1) et le deuxième vecteur propre $x^{(2)}$, écrivons le système d'équations correspondant:

$$\lambda_2 x_i^{(2)} = \sum_{i=1}^{n} a_{ij} x_i^{(2)} \quad (i = 1, 2, \ldots, n). \tag{5}$$

Eliminons des relations d'orthogonalité

$$\sum_{j=1}^{n} x_j^{(1)} x_j^{(2)} = 0 \tag{6}$$

l'une des inconnues $x_j^{(2)}$, par exemple $x_n^{(2)}$. Le système (5) sera alors remplacé par le système équivalent

$$\left.\begin{aligned} x_i^{(2)} &= \frac{1}{\lambda_2} \sum_{j=1}^{n-1} a_{ij}^{(2)} x_j^{(2)} \quad (i = 1, 2, \ldots, n-2), \\ \lambda_2 &= \frac{1}{x_{n-1}^{(2)}} \sum_{j=1}^{n-1} a_{n-1,j}^{(2)} x_j^{(2)}. \end{aligned}\right\} \tag{7}$$

En posant $x_{n-1}^{(2)} = 1$, résolvons le système (7) par la méthode itérative. On finira par obtenir la deuxième racine λ_2 de l'équation

caractéristique (1) et le vecteur propre $x^{(2)}$, la n-ième coordonnée de ce vecteur étant fournie par la condition d'orthogonalité (6). D'une façon analogue on cherche les autres racines λ_j ($j = 3, \ldots, n$) de (1) et les vecteurs propres associés $x^{(j)}$.

Nous ne considérerons pas les cas singuliers auxquels cette méthode peut donner lieu.

Exemple. Trouver pour la matrice [5]

$$A = \begin{Vmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{Vmatrix}$$

les racines λ_j de l'équation caractéristique et les vecteurs propres $x^{(j)}$.

Solution. La matrice A est symétrique et définie positive puisque

$$\Delta_1 = 4 > 0;$$

$$\Delta_2 = \begin{vmatrix} 4 & 2 \\ 2 & 5 \end{vmatrix} = 16 > 0;$$

$$\Delta_3 = \det A = 80 > 0.$$

Le système associé est de la forme

$$\left.\begin{aligned} \lambda_j x_1^{(j)} &= 4x_1^{(j)} + 2x_2^{(j)} + 2x_3^{(j)}, \\ \lambda_j x_2^{(j)} &= 2x_1^{(j)} + 5x_2^{(j)} + x_3^{(j)}, \\ \lambda_j x_3^{(j)} &= 2x_1^{(j)} + x_2^{(j)} + 6x_3^{(j)} \end{aligned}\right\} \quad (j = 1, 2, 3). \tag{8}$$

En posant $j = 1$ et $x_3^{(1)} = 1$, on obtient:

$$\left.\begin{aligned} x_1^{(1)} &= \frac{1}{\lambda_1}(4x_1^{(1)} + 2x_2^{(1)} + 2), \\ x_2^{(1)} &= \frac{1}{\lambda_1}(2x_1^{(1)} + 5x_2^{(1)} + 1), \\ \lambda_1 &= 2x_1^{(1)} + x_2^{(1)} + 6. \end{aligned}\right| \tag{9}$$

Le système (9) est résolu par la méthode itérative en prenant pour valeurs initiales

$$x_1^{(1,0)} = 1 \text{ et } x_2^{(1,0)} = 1.$$

La dernière équation du système (9) donne alors $\lambda_1^{(0)} = 9$. Les résultats du calcul sont portés sur le tableau 29.

On peut adopter

$$\lambda_1 = 8{,}3874$$

Tableau 29

Calcul par la méthode itérative des éléments propres d'une matrice relatifs à la première racine de l'équation caractéristique

k	$x_1^{(1k)}$	$x_2^{(1k)}$	$x_3^{(1k)}$	$\lambda_1^{(k)}$
0	1	1	1	9
1	0,89	0,89	1	8,67
2	0,85	0,83	1	8,53
3	0,83	0,80	1	8,46
4	0,81	0,78	1	8,40
5	0,805	0,770	1	8,38
6	0,806	0,771	1	8,383
7	0,807	0,771	1	8,385
8	0,8074	0,7715	1	8,3863
9	0,8076	0,7717	1	8,3869
10	0,8076	0,7719	1	8,3871
11	0,8077	0,7720	1	8,3874

et

$$x^{(1)} = \begin{bmatrix} 0,8077 \\ 0,7720 \\ 1 \end{bmatrix}.$$

Posons maintenant dans le système (8) $j = 2$. La condition d'orthogonalité des vecteurs $x^{(1)}$ et $x^{(2)}$ conduit à

$$0,8077\ x_1^{(2)} + 0,7720\ x_2^{(2)} + x_3^{(2)} = 0.$$

D'où

$$x_3^{(2)} = -0,8077 x_1^{(2)} - 0,7720 x_2^{(2)}. \tag{10}$$

En portant cette expression dans le système (8) et en posant $x_2^{(2)} = 1$, on obtient:

$$\left.\begin{aligned} x_1^{(2)} &= \frac{1}{\lambda_2}(2,3846 x_1^{(2)} + 0,4560), \\ \lambda_2 &= 1,1923 x_1^{(2)} + 4,2280. \end{aligned}\right\} \tag{11}$$

Le système (11) est résolu par la méthode itérative en posant

$$x_1^{(2,0)} = 1 \quad \text{et} \quad \lambda_2^{(0)} = 5,42.$$

Les résultats du calcul figurent dans le tableau 30.

On peut adopter $\lambda_2 = 4,4867$ et $x_1^{(2)} = 0,2170$; $x_2^{(2)} = 1$.

La troisième coordonnée est déterminée à partir des relations d'orthogonalité (10):

$$x_3^{(2)} = -0,9473,$$

Tableau 30

Calcul par la méthode itérative des éléments propres d'une matrice relatifs à la deuxième racine de l'équation caractéristique

k	$x_1^{(2k)}$	$x_2^{(2k)}$	$\lambda_2^{(k)}$	k	$x_1^{(2k)}$	$x_2^{(2k)}$	$\lambda_2^{(k)}$
0	1	1	5,42	6	0,223	1	4,494
1	0,52	1	4,85	7	0,220	1	4,490
2	0,35	1	4,64	8	0,218	1	4,488
3	0,28	1	4,56	9	0,2174	1	4,487
4	0,25	1	4,53	10	0,2171	1	4,4868
5	0,23	1	4,500	11	0,2170	1	4,4867

c'est pourquoi

$$x^{(2)} = \begin{bmatrix} 0{,}2170 \\ 1 \\ -0{,}9473 \end{bmatrix}.$$

Le troisième vecteur propre $x^{(3)}$ se déduit directement des deux relations d'orthogonalité

$$\left.\begin{array}{l} 0{,}8077x_1^{(3)} + 0{,}7720x_2^{(3)} + x_3^{(3)} = 0, \\ 0{,}2170x_1^{(3)} + x_2^{(3)} - 0{,}9473x_3^{(3)} = 0. \end{array}\right\}$$

En posant $x_1^{(3)} = 1$, on obtient $x_2^{(3)} = -0{,}5673$; $x_3^{(3)} = -0{,}3698$. Par conséquent,

$$x^{(3)} = \begin{bmatrix} 1 \\ -0{,}5673 \\ -0{,}3698 \end{bmatrix}.$$

La dernière équation du système (8) pour $j = 3$ conduit également à

$$\lambda_3 = 2{,}1260.$$

Pour vérifier, composons la trace de la matrice A:

$$\text{Sp}\, A = \lambda_1 + \lambda_2 + \lambda_3 = 8{,}3874 + 4{,}4867 + 2{,}1260 =$$
$$= 15{,}0001 \approx 4 + 5 + 6.$$

Remarquons que les racines fournies par le processus itératif sont le plus souvent rangées dans l'ordre décroissant de leurs modules. Les vecteurs propres de la matrice sont déterminés à un coefficient de proportionnalité près, pour cette raison, toutes les solutions du système (8) sont les suivantes:

λ_j	$x_1^{(j)}$	$x_2^{(j)}$	$x_3^{(j)}$
8,3874	$0{,}8077c_1$	$0{,}7720c_1$	c_1
4,4867	$0{,}2170c_2$	c_2	$-0{,}9473c_2$
2,1260	c_3	$-0{,}5673c_3$	$-0{,}3698c_3$

(c_1, c_2, c_3 sont des constantes arbitraires différentes du zéro).

§ 16. Inversion d'une matrice à l'aide des coefficients d'un polynôme caractéristique

Dans ce qui précède nous avons exposé les procédés de développement du déterminant caractéristique en un polynôme (§§ 3-9). En utilisant les coefficients de ce polynôme et en composant les puissances A, A^2, ..., A^{n-1} de la matrice régulière A d'ordre n il est relativement facile d'obtenir l'inverse A^{-1}. Sous ce rapport, la méthode de Leverrier (§ 8) présente de grands avantages.

Soit la matrice régulière A d'ordre n. Considérons son polynôme caractéristique

$$\det(\lambda E - A) = \lambda^n + p_1\lambda^{n-1} + \dots + p_{n-1}\lambda + p_n.$$

D'après l'identité d'Hamilton-Cayley (chapitre XI, § 2), on a

$$A^n + p_1A^{n-1} + \dots + p_{n-1}A + p_nE = 0. \tag{1}$$

En prémultipliant l'égalité matricielle (1) par A^{-1}, on obtient

$$A^{n-1} + p_1A^{n-2} + \dots + p_{n-1}E + p_nA^{-1} = 0. \tag{2}$$

D'où pour $p_n \neq 0$

$$A^{-1} = -\frac{1}{p_n}(A^{n-1} + p_1A^{n-2} + \dots + p_{n-1}E). \tag{3}$$

Ainsi, si l'on connaît les coefficients du polynôme caractéristique de A et si l'on a composé les puissances de cette matrice jusqu'à la puissance $(n-1)$ y comprise, la matrice A^{-1} se calcule sans peine d'après la formule (3).

Constatons que si $p_n = 0$ et $p_{n-1} \neq 0$, pour obtenir la formule contenant A^{-1} l'égalité vectorielle (1) doit être prémultipliée par A^{-2}, etc.

E x e m p l e. Trouver l'inverse A^{-1} de la matrice (cf. § 8, exemple)

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}.$$

S o l u t i o n. Utilisons les puissances de A déjà trouvées (§ 8):

$$A^2 = \begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix}$$

et

$$A^3 = \begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix}.$$

Le polynôme caractéristique de A s'écrivant

$$\det(\lambda A - E) = \lambda^4 - 4\lambda^3 - 40\lambda^2 - 56\lambda - 20,$$

on obtient suivant la formule (3)

$$A^{-1} = -\frac{1}{-20}\left\{\begin{bmatrix} 208 & 178 & 192 & 242 \\ 178 & 148 & 154 & 192 \\ 192 & 154 & 148 & 178 \\ 242 & 192 & 178 & 208 \end{bmatrix} - \right.$$

$$\left. -4\begin{bmatrix} 30 & 22 & 18 & 20 \\ 22 & 18 & 16 & 18 \\ 18 & 16 & 18 & 22 \\ 20 & 18 & 22 & 30 \end{bmatrix} - 40\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} - 56\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right\} =$$

$$= \frac{1}{10}\left\{\begin{bmatrix} 104 & 89 & 96 & 121 \\ 89 & 74 & 77 & 96 \\ 96 & 77 & 74 & 89 \\ 121 & 96 & 89 & 104 \end{bmatrix} - \begin{bmatrix} 60 & 44 & 36 & 40 \\ 44 & 36 & 32 & 36 \\ 36 & 32 & 36 & 44 \\ 40 & 36 & 44 & 60 \end{bmatrix} - \right.$$

$$\left. - \begin{bmatrix} 20 & 40 & 60 & 80 \\ 40 & 20 & 40 & 60 \\ 60 & 40 & 20 & 40 \\ 80 & 60 & 40 & 20 \end{bmatrix} - \begin{bmatrix} 28 & 0 & 0 & 0 \\ 0 & 28 & 0 & 0 \\ 0 & 0 & 28 & 0 \\ 0 & 0 & 0 & 28 \end{bmatrix}\right\} =$$

$$= \begin{bmatrix} -0{,}4 & 0{,}5 & 0 & 0{,}1 \\ 0{,}5 & -1 & 0{,}5 & 0 \\ 0 & 0{,}5 & -1 & 0{,}5 \\ 0{,}1 & 0 & 0{,}5 & -0{,}4 \end{bmatrix}.$$

Pour vérifier, composons le produit

$$AA^{-1} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} -0{,}4 & 0{,}5 & 0 & 0{,}1 \\ 0{,}5 & -1 & 0{,}5 & 0 \\ 0 & 0{,}5 & -1 & 0{,}5 \\ 0{,}1 & 0 & 0{,}5 & -0{,}4 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = E.$$

§ 17. Méthode de Lusternik pour améliorer la convergence du processus itératif de résolution d'un système d'équations linéaires

Supposons que le système d'équations linéaires

$$Ax = b \tag{1}$$

est réduit à la forme commode pour l'itération

$$x = \beta + \alpha x. \tag{1'}$$

D'après la méthode itérative (chapitre IV, § 8), les approximations successives de la solution x du système (1') sont définies par la formule

$$x^{(m)} = \beta + \alpha x^{(m-1)} \qquad (m = 1, 2, \ldots), \tag{2}$$

où $x^{(0)}$ est un vecteur initial arbitraire.

Supposons que les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ de la matrice α sont distinctes et que

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|. \tag{3}$$

Le processus itératif (2) converge si

$$|\lambda_1| < 1.$$

La première valeur propre λ_1 peut être déterminée approximativement moyennant les méthodes décrites aux §§ 11 et 12. Comme l'a démontré L. Lusternik [6], en utilisant le nombre λ_1 on peut accélérer nettement la convergence du processus itératif (2) de la résolution du système (1').

Si m est suffisamment grand, on peut poser approximativement

$$x \approx x^{(m)}.$$

Evaluons l'erreur $x - x^{(m)}$. Sous la condition de convergence du processus (2), on a

$$x = \lim_{m\to\infty} x^{(m)} = x^{(0)} + \sum_{k=1}^{m} (x^{(k)} - x^{(k-1)});$$

d'autre part,

$$x^{(m)} = x^{(0)} + \sum_{k=1}^{m} (x^{(k)} - x^{(k-1)}).$$

Donc

$$x - x^{(m)} = \sum_{k=m+1}^{\infty} (x^{(k)} - x^{(k-1)}) = \\ = [x^{(m+1)} - x^{(m)}] + [x^{(m+2)} - x^{(m+1)}] + \ldots \quad (4)$$

Puisque

$$x^{(k)} - x^{(k-1)} = [\beta + \alpha x^{(k-1)}] - [\beta + \alpha x^{(k-2)}] = \\ = \alpha (x^{(k-1)} - x^{(k-2)}) = \alpha^{k-1} (x^{(1)} - x^{(0)}) \quad \text{avec } k = 1, 2, \ldots,$$

il vient

$$x - x^{(m)} = \alpha^m (x^{(1)} - x^{(0)}) + \alpha^{m+1} (x^{(1)} - x^{(0)}) + \ldots \quad (5)$$

Soient $y_1, y_2, \ldots, y_n$ les vecteurs propres de la matrice α associés aux valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ et formant la base de l'espace E_n. En développant le vecteur $x^{(1)} - x^{(0)}$ par rapport aux vecteurs de cette base, on aura:

$$x^{(1)} - x^{(0)} = c_1 y_1 + c_2 y_2 + \ldots + c_n y_n,$$

où c_j $(j = 1, 2, \ldots, n)$ sont certains nombres déterminés. On en tire

$$x^{(k)} - x^{(k-1)} = \alpha^{k-1} (x^{(1)} - x^{(0)}) = \\ = c_1 \lambda_1^{k-1} y_1 + c_2 \lambda_2^{k-1} y_2 + \ldots + c_n \lambda_n^{k-1} y_n \quad (6) \\ (k = m+1, m+2, \ldots).$$

On obtient donc d'après la formule (5):

$$x - x^{(m)} = c_1 \lambda_1^m (1 + \lambda_1 + \lambda_1^2 + \ldots) y_1 + \\ + c_2 \lambda_2^m (1 + \lambda_2 + \lambda_2^2 + \ldots) y_2 + \ldots + c_n \lambda_n^m (1 + \lambda_n + \lambda_n^2 + \ldots) y_n = \\ = \frac{c_1 \lambda_1^m}{1 - \lambda_1} y_1 + \frac{c_2 \lambda_2^m}{1 - \lambda_2} y_2 + \ldots + \frac{c_n \lambda_n^m}{1 - \lambda_n} y_n.$$

D'où, compte tenu de l'inégalité (3),

$$x - x^{(m)} = \frac{c_1 \lambda_1^m}{1 - \lambda_1} y_1 + O(\lambda_2^m). \quad (7)$$

De plus, on déduit de la formule (6) pour $k = m + 1$:

$$x^{(m+1)} - x^{(m)} = c_1 \lambda_1^m y_1 + O(\lambda_2^m). \quad (8)$$

Par suite,

$$x - x^{(m)} = \frac{x^{(m+1)} - x^{(m)}}{1-\lambda_1} + O(\lambda_2^m).$$

Ainsi, on a finalement:

$$x \approx x^{(m)} + \frac{x^{(m+1)} - x^{(m)}}{1-\lambda_1}. \tag{9}$$

Le terme supplémentaire $\frac{x^{(m+1)} - x^{(m)}}{1-\lambda_1}$ améliore sensiblement la convergence du processus itératif (2).

Comme la formule (8) entraîne

$$x^{(m+1)} - x^{(m)} = \lambda_1 (x^{(m)} - x^{(m-1)}) + O(\lambda_2^m), \tag{10}$$

la formule (9) peut être remplacée par la formule suivante

$$x \approx x^{(m)} + \frac{\lambda_1}{1-\lambda_1}(x^{(m)} - x^{(m-1)}). \tag{11}$$

La formule (11) rend inutile le calcul de l'approximation successive.

En vertu de la formule (10), la plus grande valeur propre λ_1 peut être calculée d'après la formule

$$\lambda_1 \approx \frac{(x^{(m)} - x^{(m-1)})_i}{(x^{(m-1)} - x^{(m-2)})_i} \qquad (i = 1, 2, \ldots, n).$$

Dans le cas d'une matrice α symétrique, en utilisant la méthode des produits scalaires on obtient une formule plus précise:

$$\lambda_1 \approx \frac{(x^{(m)} - x^{(m-1)},\ x^{(m)} - x^{(m-1)})}{(x^{(m-1)} - x^{(m-2)},\ x^{(m)} - x^{(m-1)})}.$$

En particulier, si

$$x^{(0)} = \beta,$$

il vient

$$x^{(m)} - x^{(m-1)} = \alpha^{m-1}(x^{(1)} - x^{(0)}) = \alpha^m \beta$$

et

$$x^{(m)} = x^{(0)} + \sum_{k=1}^{m} \alpha^{k-1}(x^{(1)} - x^{(0)}) = \sum_{k=0}^{m} \alpha^k \beta.$$

On a donc

$$\lambda_1 \approx \frac{(\alpha^m \beta)_i}{(\alpha^{m-1}\beta)_i} \qquad (i = 1, 2, \ldots, n), \tag{12}$$

où $(\alpha^m \beta)_i$ et $(\alpha^{m-1}\beta)_i$ sont les i-èmes coordonnées respectivement des vecteurs $\alpha^m \beta$ et $\alpha^{m-1}\beta$. D'une façon analogue, si la matrice α est symétrique

$$\lambda_1 \approx \frac{(\alpha^m \beta,\ \alpha^m \beta)}{(\alpha^{m-1}\beta,\ \alpha^m \beta)}. \tag{13}$$

Exemple. Résoudre le système suivant par itération [1]

$$\left.\begin{aligned} 0{,}78x_1-0{,}02x_2-0{,}12x_3-0{,}14x_4&=0{,}76;\\ -0{,}02x_1+0{,}86x_2-0{,}04x_3+0{,}06x_4&=0{,}08;\\ -0{,}12x_1-0{,}04x_2+0{,}72x_3-0{,}08x_4&=1{,}12;\\ -0{,}14x_1+0{,}06x_2-0{,}08x_3+0{,}74x_4&=0{,}68,\end{aligned}\right\}$$

en appliquant pour améliorer la précision de la solution la méthode de Lusternik.

Solution. Réduisons le système à la forme commode pour l'application de la méthode itérative

$$\left.\begin{aligned} x_1&=0{,}22x_1+0{,}02x_2+0{,}12x_3+0{,}14x_4+0{,}76;\\ x_2&=0{,}02x_1+0{,}14x_2+0{,}04x_3-0{,}06x_4+0{,}08;\\ x_3&=0{,}12x_1+0{,}04x_2+0{,}28x_3+0{,}08x_4+1{,}12;\\ x_4&=0{,}14x_1-0{,}06x_2+0{,}08x_3+0{,}26x_4+0{,}68\end{aligned}\right\} \qquad (14)$$

ou en le mettant sous une forme matricielle

$$\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\end{bmatrix}=\begin{bmatrix}0{,}76\\ 0{,}08\\ 1{,}12\\ 0{,}68\end{bmatrix}+\begin{bmatrix}0{,}22 & 0{,}02 & 0{,}12 & 0{,}14\\ 0{,}02 & 0{,}14 & 0{,}04 & -0{,}06\\ 0{,}12 & 0{,}04 & 0{,}28 & 0{,}08\\ 0{,}14 & -0{,}06 & 0{,}08 & 0{,}26\end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\end{bmatrix}. \qquad (14')$$

D'où

$$\alpha=\begin{bmatrix}0{,}22 & 0{,}02 & 0{,}12 & 0{,}14\\ 0{,}02 & 0{,}14 & 0{,}04 & -0{,}06\\ 0{,}12 & 0{,}04 & 0{,}28 & 0{,}08\\ 0{,}14 & -0{,}06 & 0{,}08 & 0{,}26\end{bmatrix} \quad \text{et} \quad \beta=\begin{bmatrix}0{,}76\\ 0{,}08\\ 1{,}12\\ 0{,}68\end{bmatrix}.$$

Comme

$$\|\alpha\|_m=\max(0{,}50;\ 0{,}26;\ 0{,}52;\ 0{,}54)=0{,}54<1,$$

le processus itératif de (14) est convergent.

En prenant le vecteur β pour le vecteur initial $x^{(0)}$ on obtient pour la m-ième approximation $x^{(m)}$ de la solution cherchée

$$x=\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\end{bmatrix}$$

l'expression suivante:

$$x^{(m)}=\sum_{k=0}^{m}\alpha^k\beta. \qquad (15)$$

Ainsi, pour calculer $x^{(m)}$, il faut former les itérations successives du vecteur β à l'aide de la matrice α. On a:

$$\alpha\beta = \begin{bmatrix} 0,22 & 0,02 & 0,12 & 0,14 \\ 0,02 & 0,14 & 0,04 & -0,06 \\ 0,12 & 0,04 & 0,28 & 0,08 \\ 0,14 & -0,06 & 0,08 & 0,26 \end{bmatrix} \begin{bmatrix} 0,76 \\ 0,08 \\ 1,12 \\ 0,68 \end{bmatrix} = \begin{bmatrix} 0,3984 \\ 0,0304 \\ 0,4624 \\ 0,3680 \end{bmatrix};$$

$$\alpha^2\beta = \alpha\cdot\alpha\beta = \begin{bmatrix} 0,22 & 0,02 & 0,12 & 0,14 \\ 0,02 & 0,14 & 0,04 & -0,06 \\ 0,12 & 0,04 & 0,28 & 0,08 \\ 0,14 & -0,06 & 0,08 & 0,26 \end{bmatrix} \begin{bmatrix} 0,3984 \\ 0,0304 \\ 0,4624 \\ 0,3680 \end{bmatrix} =$$

$$= \begin{bmatrix} 0,195264 \\ 0,008640 \\ 0,207936 \\ 0,186624 \end{bmatrix},$$

etc.

Les resultats des calculs correspondants sont donnés dans le tableau 31.

Tableau 31

Itérations successives du vecteur β par la matrice α

β	$\alpha\beta$	$\alpha^2\beta$	$\alpha^3\beta$	$\alpha^4\beta$
0,76	0,3984	0,195264	0,09421056	0,04527913
0,08	0,0304	0,008640	0,00223488	0,00055572
1,12	0,4624	0,207936	0,09692928	0,04589292
0,68	0,3680	0,186624	0,09197568	0,04472340

$\alpha^5\beta$	$\alpha^6\beta$	$\alpha^7\beta$	$\alpha^8\beta$	$x^{(8)} = \sum_{k=0}^{8} \alpha^k\beta$
0,02174095	0,01043649	0,00500961	0,00240463	1,532746
0,00013570	0,00003285	0,00000792	0,00000190	0,122009
0,02188361	0,01047017	0,00501763	0,00240654	1,972937
0,02160525	0,01040364	0,00500170	0,00240272	1,410737

Dans la formule (11) adoptons $m = 8$. La matrice α étant symétrique, utilisons pour le calcul de sa première valeur propre λ_1 la

méthode des produits scalaires. On a :

$$\lambda_1 \approx \frac{(\alpha^8\beta,\ \alpha^8\beta)}{(\alpha^7\beta,\ \alpha^8\beta)} =$$

$$= \frac{240\,463^2 + 190^2 + 240\,654^2 + 240\,272^2}{500\,961 \cdot 240\,463 + 792 \cdot 190 + 501\,763 \cdot 240\,654 + 500\,170 \cdot 240\,272} = 0{,}480000.$$

On en tire, compte tenu de $x^{(8)} - x^{(7)} = \alpha^8\beta$,

$$x \approx x^{(8)} + \lambda_1 \frac{\alpha^8\beta}{1-\lambda_1} =$$

$$= \begin{bmatrix} 1{,}532746 \\ 0{,}122009 \\ 1{,}972937 \\ 1{,}410737 \end{bmatrix} + \frac{12}{13} \begin{bmatrix} 0{,}002405 \\ 0{,}000002 \\ 0{,}002406 \\ 0{,}002403 \end{bmatrix} = \begin{bmatrix} 1{,}534965 \\ 0{,}122011 \\ 1{,}975159 \\ 1{,}412955 \end{bmatrix}.$$

Voici à titre de comparaison les valeurs de la solution du système (11) fournies par la méthode de Gauss [1]:

$$x_1 = 1{,}534965; \quad x_2 = 0{,}122010;$$
$$x_3 = 1{,}975166; \quad x_4 = 1{,}412955.$$

Ainsi, si $x^{(8)}$ donnait les valeurs de x_i ($i = 1, 2, 3, 4$) avec une précision de $1 \cdot 10^{-3}$ à $2 \cdot 10^{-3}$, après les corrections de Lusternik, la précision sera poussée à peu près à 10^{-6}.

La méthode de Lusternik peut être appliquée également au processus de Seidel. On sait que pour le système (2) le processus de Seidel est un procédé itératif d'un système équivalent

$$x = \beta_1 + \alpha_1 x,$$

où la matrice α_1 est définie par la matrice α (cf. chapitre XI, § 3), et notamment, si

$$\alpha = B + C,$$

où B est une matrice triangulaire inférieure à diagonale nulle et C une matrice triangulaire supérieure, alors

$$\alpha_1 = (E - B)^{-1} C.$$

Donc si $\xi^{(m)}$ ($m = 1, 2, \ldots$) sont des approximations successives de la solution x du système (2) établies d'après Seidel, on peut poser :

$$x \approx \xi^{(m)} + \frac{\xi^{(m+1)} - \xi^{(m)}}{1-\mu_1},$$

avec μ_1 la valeur propre de α_1 la plus grande en module.

Notons qu'il existe également d'autres méthodes d'amélioration de la convergence des processus itératifs de résolution des systèmes d'équations linéaires, celles, entre autres, de M. Gavourine [7], [8], de A. Abramov.

BIBLIOGRAPHIE

1. *V. Faddéeva*. Méthodes numériques d'algèbre linéaire. Gostekhizdat, Moscou, 1950, chapitre III.
2. *I. Guelfand*. Cours d'algèbre linéaire. Ed. 2. Gostekhizdat, Moscou-Léningrad, 1951, appendice I.
3. *A. Kurosh*. Cours d'algèbre supérieure. Editions Mir, Moscou, 1971.
4. *H. Wailend*. Représentation de l'équation caractéristique sous la forme d'un polynôme. Succès des sciences mathématiques II, fascicule 4 (20) (1947), 128-158.
5. *W. E. Milne*. Numerical calculus. Princeton University Press. Princeton, 1949.
6. *L. Lusternik*. Travaux de l'Institut des mathématiques V. Steklov, 20 (1947), p. 49.
7. *M. Gavourine*. Application des polynômes de meilleure approximation à l'amélioration de la convergence des processus itératifs. Succès des sciences mathématiques 5; 3 (37) (1950), 156-160.
8. *I. Bérézine*, *N. Jidkov*. Méthodes de calcul. Fizmatguiz, 1959, t. 2, chapitre VIII.

CHAPITRE XIII

RÉSOLUTION APPROCHÉE DES SYSTÈMES D'ÉQUATIONS NON LINÉAIRES

§ 1. Méthode de Newton

Considérons un système d'équations en général non linéaire

$$\left.\begin{aligned} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{aligned}\right\} \tag{1}$$

à premiers membres réels.

Ecrivons le système (1) sous une forme abrégée. L'ensemble des arguments $x_1, x_2, \ldots, x_n$ peut être considéré comme un vecteur de dimension n

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

De façon analogue, l'ensemble des fonctions $f_1, f_2, \ldots, f_n$ forme un vecteur de dimension n (*vecteur fonction*)

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}.$$

Le système (1) peut donc s'écrire sous une forme abrégée

$$f(x) = 0. \tag{1'}$$

Pour résoudre le système (1') on fera appel à la méthode des approximations successives.

Supposons qu'on ait trouvé la p-ième approximation

$$x^{(p)} = (x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})$$

d'une des solutions isolées $x = (x_1, x_2, \ldots, x_n)$ de l'équation vectorielle (1'). La solution exacte de (1') pourra alors se mettre

sous la forme

$$x = x^{(p)} + \varepsilon^{(p)}, \tag{2}$$

où $\varepsilon^{(p)} = (\varepsilon_1^{(p)}, \varepsilon_2^{(p)}, \ldots, \varepsilon_n^{(p)})$ est une correction (*erreur de solution*).

En portant l'expression (2) dans (1′), on aura

$$f(x^{(p)} + \varepsilon^{(p)}) = 0. \tag{3}$$

Supposons que la fonction $f(x)$ soit continûment dérivable dans un certain domaine convexe qui contient x et $x^{(p)}$ et décomposons le premier membre de l'équation (3) par rapport aux puissances du petit vecteur $\varepsilon^{(p)}$ en nous bornant aux termes linéaires

$$f(x^{(p)} + \varepsilon^{(p)}) = f(x^{(p)}) + f'(x^{(p)})\,\varepsilon^{(p)} = 0 \tag{4}$$

ou, sous une forme développée,

$$\left.\begin{aligned}
&f_1(x_1^{(p)} + \varepsilon_1^{(p)},\ x_2^{(p)} + \varepsilon_2^{(p)},\ \ldots,\ x_n^{(p)} + \varepsilon_n^{(p)}) = \\
&\quad = f_1(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{1x_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} + \\
&\qquad + f'_{1x_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots \\
&\qquad \ldots + f'_{1x_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0, \\
&f_2(x_1^{(p)} + \varepsilon_1^{(p)},\ x_2^{(p)} + \varepsilon_2^{(p)},\ \ldots,\ x_n^{(p)} + \varepsilon_n^{(p)}) = \\
&\quad = f_2(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{2x_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} + \\
&\qquad + f'_{2x_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots \\
&\qquad \ldots + f'_{2x_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&f_n(x_1^{(p)} + \varepsilon_1^{(p)},\ x_2^{(p)} + \varepsilon_2^{(p)},\ \ldots,\ x_n^{(p)} + \varepsilon_n^{(p)}) = \\
&\quad = f_n(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}) + f'_{nx_1}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_1^{(p)} + \\
&\qquad + f'_{nx_2}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_2^{(p)} + \ldots \\
&\qquad \ldots + f'_{nx_n}(x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)})\,\varepsilon_n^{(p)} = 0.
\end{aligned}\right\} \tag{4'}$$

Les formules (4) et (4′) entraînent qu'il faut entendre par dérivée $f'(x)$ la *matrice jacobienne* du système des fonctions $f_1, f_2, \ldots, f_n$ des variables $x_1, x_2, \ldots, x_n$

$$f'(x) = W(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \ldots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \ldots & \frac{\partial f_2}{\partial x_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \ldots & \frac{\partial f_n}{\partial x_n} \end{bmatrix},$$

ou, en écriture condensée,

$$f'(x) = W(x) = \left[\frac{\partial f_i}{\partial x_j}\right] \qquad (i, j = 1, 2, \ldots, n).$$

(4′) est un système linéaire par rapport aux erreurs $\varepsilon_i^{(p)}$ ($i = 1, 2, \ldots, n$) à matrice $W(x)$; aussi peut-on mettre la formule (4) sous la forme:

$$f(x^{(p)}) + W(x^{(p)})\,\varepsilon^{(p)} = 0.$$

En supposant que la matrice $W(x^{(p)})$ est régulière, on obtient:

$$\varepsilon^{(p)} = -W^{-1}(x^{(p)})\,f(x^{(p)}).$$

Par conséquent,

$$x^{(p+1)} = x^{(p)} - W^{-1}(x^{(p)})\,f(x^{(p)}) \qquad (p = 0, 1, 2, \ldots) \tag{5}$$

(*méthode de Newton*).

On prend pour approximation initiale $x^{(0)}$ une valeur grossière de la solution cherchée.

Exemple 1. Trouver les solutions positives approchées du système d'équations (cf. chapitre IV, § 9)

$$\left.\begin{aligned} f_1(x_1, x_2) &\equiv x_1 + 3\lg x_1 - x_2^2 = 0, \\ f_2(x_1, x_2) &\equiv 2x_1^2 - x_1x_2 - 5x_1 + 1 = 0. \end{aligned}\right\} \tag{6}$$

Solution. Les courbes définies par le système (6) se coupent approximativement aux points M_1 (1,4; —1,5) et M_2 (3,4; 2,2). En partant de l'approximation initiale

$$x^{(0)} = \begin{bmatrix} 3{,}4 \\ 2{,}2 \end{bmatrix},$$

calculons les deuxièmes approximations des solutions, en effectuant les calculs avec quatre décimales. En posant

$$f(x) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix},$$

on a

$$f(x^{(0)}) = \begin{bmatrix} 3{,}4 + 3\lg 3{,}4 - 2{,}2^2 \\ 2\cdot 3{,}4^2 - 3{,}4\cdot 2{,}2 - 5\cdot 3{,}4 + 1 \end{bmatrix} = \begin{bmatrix} 0{,}1544 \\ -0{,}3600 \end{bmatrix}.$$

Composons la matrice jacobienne

$$W(x) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 + \dfrac{3M}{x_1} & -2x_2 \\ 4x_1 - x_2 - 5 & -x_1 \end{bmatrix},$$

avec $M = 0{,}43429$. D'où

$$W(x^{(0)}) = \begin{bmatrix} 1 + \frac{3\cdot 0{,}43429}{3{,}4} & -2\cdot 2{,}2 \\ 4\cdot 3{,}4 - 2{,}2 - 5 & -3{,}4 \end{bmatrix} = \begin{bmatrix} 1{,}3832 & -4{,}4 \\ 6{,}4 & -3{,}4 \end{bmatrix},$$

en outre

$$\Delta = \det W(x^{(0)}) = 23{,}4571.$$

La matrice $W(x^{(0)})$ est donc une matrice régulière. Composons son inverse

$$W^{-1}(x^{(0)}) = \frac{1}{\Delta} \begin{bmatrix} -3{,}4 & 6{,}4 \\ -4{,}4 & 1{,}3832 \end{bmatrix}.$$

La formule (5) donne

$$x^{(1)} = \begin{bmatrix} 3{,}4 \\ 2{,}2 \end{bmatrix} - \frac{1}{23{,}4571} \begin{bmatrix} -3{,}4 & 6{,}4 \\ -4{,}4 & 1{,}3832 \end{bmatrix} \begin{bmatrix} 0{,}1544 \\ -0{,}3600 \end{bmatrix} =$$

$$= \begin{bmatrix} 3{,}4 \\ 2{,}2 \end{bmatrix} - \frac{1}{23{,}4571} \begin{bmatrix} -2{,}10896 \\ -1{,}48604 \end{bmatrix} = \begin{bmatrix} 3{,}4 \\ 2{,}2 \end{bmatrix} + \begin{bmatrix} 0{,}0899 \\ 0{,}0633 \end{bmatrix} = \begin{bmatrix} 3{,}4899 \\ 2{,}2633 \end{bmatrix}.$$

Les approximations ultérieures s'obtiennent d'une façon analogue. Les résultats du calcul sont fournis par le tableau 32.

Tableau 32

Approximations successives des solutions du système (6)

i	x_1	$\varepsilon_1 = \Delta x_1$	x_2	$\varepsilon_2 = \Delta x_2$
0	3,4	0,0899	2,2	0,0633
1	3,4899	−0,0008	2,2633	−0,0012
2	3,4891	−0,0016	2,2621	−0,0005
3	3,4875		2,2616	

Si l'on s'arrête à l'approximation $x^{(3)}$, on a:

$$x_1 = 3{,}4875; \quad x_2 = 2{,}2616,$$

et

$$f(x^{(3)}) = \begin{bmatrix} 0{,}0002 \\ 0{,}0000 \end{bmatrix}.$$

Exemple 2. Trouver par la méthode de Newton la solution positive approchée du système d'équations

$$\left.\begin{aligned} x^2 + y^2 + z^2 &= 1, \\ 2x^2 + y^2 - 4z &= 0, \\ 3x^2 - 4y + z^2 &= 0, \end{aligned}\right\}$$

en partant de l'approximation initiale

$$x_0 = y_0 = z_0 = 0{,}5.$$

Solution. On a

$$f(x) = \begin{bmatrix} x^2+y^2+z^2-1 \\ 2x^2+y^2-4z \\ 3x^2-4y+z^2 \end{bmatrix}.$$

D'où

$$f(x^{(0)}) = \begin{bmatrix} 0{,}25+0{,}25+0{,}25-1 \\ 0{,}50+0{,}25-2{,}00 \\ 0{,}75-2{,}00+0{,}25 \end{bmatrix} = \begin{bmatrix} -0{,}25 \\ -1{,}25 \\ -1{,}00 \end{bmatrix}.$$

Formons la matrice jacobienne

$$W(x) = \begin{bmatrix} 2x & 2y & 2z \\ 4x & 2y & -4 \\ 6x & -4 & 2z \end{bmatrix}.$$

On a

$$W(x^{(0)}) = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{bmatrix}$$

et

$$\det W(x^{(0)}) = \begin{vmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{vmatrix} = -40.$$

Cherchons la matrice inverse

$$W^{-1}(x^{(0)}) = -\frac{1}{40}\begin{bmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{bmatrix} = \begin{bmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & -\frac{3}{20} \\ \frac{11}{40} & -\frac{7}{40} & \frac{1}{40} \end{bmatrix}.$$

D'après la formule (5) la première approximation est

$$x^{(1)} = x^{(0)} - W^{-1}(x^{(0)}) f(x^{(0)}) =$$

$$= \begin{bmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{bmatrix} - \begin{bmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{7}{20} & \frac{1}{20} & \frac{-3}{20} \\ \frac{11}{40} & \frac{-7}{40} & \frac{1}{40} \end{bmatrix} \begin{bmatrix} -0{,}25 \\ -1{,}25 \\ -1{,}00 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{bmatrix} + \begin{bmatrix} 0{,}375 \\ 0 \\ -0{,}125 \end{bmatrix} = \begin{bmatrix} 0{,}875 \\ 0{,}500 \\ 0{,}375 \end{bmatrix}.$$

Calculons ensuite la deuxième approximation $x^{(2)}$. On a

$$f(x^{(1)}) = \begin{bmatrix} 0{,}875^2 + 0{,}500^2 + 0{,}375^2 - 1 \\ 2\cdot 0{,}875^2 + 0{,}500^2 - 4\cdot 0{,}375 \\ 3\cdot 0{,}875^2 - 4\cdot 0{,}500 + 0{,}375^2 \end{bmatrix} = \begin{bmatrix} 0{,}15625 \\ 0{,}28125 \\ 0{,}43750 \end{bmatrix}$$

et

$$W(x^{(1)}) = \begin{bmatrix} 2\cdot 0{,}875 & 2\cdot 0{,}500 & 2\cdot 0{,}375 \\ 4\cdot 0{,}875 & 2\cdot 0{,}500 & -4 \\ 6\cdot 0{,}875 & -4 & 2\cdot 0{,}375 \end{bmatrix} = \begin{bmatrix} 1{,}750 & 1 & 0{,}750 \\ 3{,}500 & 1 & -4 \\ 5{,}250 & -4 & 0{,}750 \end{bmatrix}.$$

D'où

$$\det W(x^{(1)}) = \begin{vmatrix} 1{,}750 & 1 & 0{,}750 \\ 3{,}500 & 1 & -4 \\ 5{,}250 & -4 & 0{,}750 \end{vmatrix} = \begin{vmatrix} 1{,}750 & 1 & 0{,}750 \\ 1{,}750 & 0 & -4{,}750 \\ 12{,}250 & 0 & 3{,}750 \end{vmatrix} = -64{,}75$$

et

$$W^{-1}(x^{(1)}) = -\frac{1}{64{,}75} \begin{bmatrix} -15{,}25 & -3{,}75 & -4{,}75 \\ -23{,}625 & -2{,}6250 & 9{,}625 \\ -19{,}25 & 12{,}25 & -1{,}75 \end{bmatrix}.$$

En appliquant la formule (5), on obtient :

$$x^{(2)} = x^{(1)} - W^{-1}(x^{(1)}) f(x^{(1)}) =$$

$$= \begin{bmatrix} 0{,}875 \\ 0{,}500 \\ 0{,}375 \end{bmatrix} + \frac{1}{64{,}75} \begin{bmatrix} -15{,}25 & -3{,}75 & -4{,}75 \\ -23{,}625 & -2{,}6250 & 9{,}625 \\ -19{,}25 & 12{,}25 & -1{,}75 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 0{,}15625 \\ 0{,}28125 \\ 0{,}43750 \end{bmatrix} = \begin{bmatrix} 0{,}875 \\ 0{,}500 \\ 0{,}375 \end{bmatrix} - \begin{bmatrix} 0{,}08519 \\ 0{,}00338 \\ 0{,}00507 \end{bmatrix} = \begin{bmatrix} 0{,}78981 \\ 0{,}49662 \\ 0{,}36993 \end{bmatrix}.$$

De façon analogue on calcule les approximations suivantes :

$$x^{(3)} = \begin{bmatrix} 0{,}78521 \\ 0{,}49662 \\ 0{,}36992 \end{bmatrix}, \quad f(x^{(3)}) = \begin{bmatrix} 0{,}00001 \\ 0{,}00004 \\ 0{,}00005 \end{bmatrix},$$

etc.

En se bornant à la troisième approximation, on a

$$x = 0{,}7852\,; \quad y = 0{,}4966\,; \quad z = 0{,}3699.$$

§ 2. Remarques générales sur la convergence du processus de Newton

Le § 1 donne un exposé formel de la méthode de Newton. Les conditions de convergence de cette méthode dans le cas d'un système ont été étudiées par Willers, Sténine, Ostrowski, Kantorovitch, d'autres encore. Nous exposons dans ce qui suit un cas particulier relatif aux systèmes finis d'équations non linéaires du théorème de Kantorovitch (théorème 1) [1] sur la convergence du processus de Newton dans des espaces fonctionnels ; pour simplifier le raisonnement, on utilise des estimations plus grossières. D'après Kantorovitch, on établit également la rapidité avec laquelle le processus de Newton converge, l'unicité d'une solution du système et la stabilité du processus par rapport au choix de l'approximation initiale (théorèmes 2 à 4). On obtient comme cas particulier le théorème d'Ostrowski [2] sur la convergence du processus de Newton pour une équation au second membre analytique complexe.

Dans ce qui suit il serait commode de considérer les ensembles des fonctions comme *vecteur fonction* ou *fonction matricielle.* Pour alléger l'exposé nous allons généraliser à ces cas la notion de la dérivée.

Soient $x = (x_1, \ldots, x_n)$ et

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix},$$

où $f_i \in C^{(1)}$ $(i = 1, 2, \ldots, n)$.

Définition 1. Par dérivée $f'(x)$ on entend la matrice jacobienne du système des fonctions f_i $(i = 1, \ldots, n)$ par rapport aux variables $x_1, \ldots, x_n$

$$f'(x) = \left[\frac{\partial f_i}{\partial x_j}\right]. \tag{1}$$

La fonction matricielle

$$F(x) = \begin{bmatrix} f_{11}(x) \ldots f_{1r}(x) \\ \ldots\ldots\ldots\ldots \\ f_{n1}(x) \ldots f_{nr}(x) \end{bmatrix}$$

peut être considérée comme un ensemble de m vecteurs fonctions

$$F_1(x) = \begin{bmatrix} f_{11}(x) \\ \vdots \\ f_{n1}(x) \end{bmatrix}, \ldots, F_r(x) = \begin{bmatrix} f_{1r}(x) \\ \vdots \\ f_{nr}(x) \end{bmatrix}.$$

Il est donc naturel d'entendre par dérivée $F'(x)$ l'ensemble

$$F'(x) = [F_1'(x) \ldots F_r'(x)],$$

ù

$$F'_k(x) = \begin{bmatrix} \frac{\partial f_{1k}}{\partial x_1} & \dots & \frac{\partial f_{1k}}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial f_{nk}}{\partial x_1} & \dots & \frac{\partial f_{nk}}{\partial x_n} \end{bmatrix}$$

sont les matrices jacobiennes ($k = 1, 2, \dots, r$).

Définition 2. Si $F(x) = [f_{ij}(x)]$ est une matrice fonctionnelle $n \times r$ et $f_{ij}(x) \in C^{(1)}$, on pose

$$F'(x) = [F'_k(x)], \tag{2}$$

où

$$F'_k(x) = \left[\frac{\partial f_{ik}}{\partial x_j}\right] \quad (i, j = 1, 2, \dots, n;\ k = 1, 2, \dots, r).$$

En particulier, si le vecteur fonction $f(x) = [f_i(x)]$ est tel que $f_i(x) \in C^{(2)}$,

$$f''(x) = [W_1(x) \dots W_n(x)],$$

avec

$$W_k(x) = \left[\frac{\partial^2 f_i}{\partial x_k\, \partial x_j}\right] \quad (k = 1, 2, \dots, n).$$

Pour évaluer les matrices nous utiliserons dans ce paragraphe la m-norme (chapitre VII, § 7) en omettant l'indice m pour abréger l'écriture:

$$\| f(x) \| = \max_i | f_i(x) |;$$

$$\| f'(x) \| = \max_i \sum_{j=1}^{n} \left| \frac{\partial f_i(x)}{\partial x_j} \right|;$$

$$\| f''(x) \| = \max_k \| W_k(x) \| = \max_k \left\{ \max_i \sum_{j=1}^{n} \left| \frac{\partial^2 f_i(x)}{\partial x_k\, \partial x_j} \right| \right\}, \quad \text{etc.}$$

D'une façon analogue

$$\| F(x) \| = \max_i \sum_{j=1}^{n} | f_{ij}(x) |,$$

$$\| F'(x) \| = \max_{i, j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(x)}{\partial x_k} \right|^{*}.$$

Déduisons au préalable quelques estimations des m-normes des différences de valeurs des fonctions matricielles analogues à la

---

* Puisqu'on a évidemment pour tout ensemble fini des nombres $\{a_{ij}\}$

$$\max_i (\max_j a_{ij}) = \max_{i, j} a_{ij}.$$

formule des accroissements finis, qui nous seront utiles dans ce qui suit (cf. [1]).

L e m m e 1. *Si*

$$F(x) = [f_{ij}(x)] \qquad (n \times r),$$

où $f_{ij}(x)$ *sont continues avec leurs dérivées premières partielles dans un domaine convexe qui contient les points* x *et* $x + \Delta x$, *alors*

$$\|F(x + \Delta x) - F(x)\| \leqslant r \|\Delta x\| \cdot \|F'(\xi)\|, \tag{3}$$

où $\xi = x + \theta \Delta x$, $0 < \theta < 1$, *et par norme des matrices on entend la m-norme.*

D é m o n s t r a t i o n. En appliquant la formule de Taylor, on obtient :

$$F(x + \Delta x) - F(x) = [f_{ij}(x + \Delta x) - f_{ij}(x)] = \left[\sum_{k=1}^{n} \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \Delta x_k\right]$$

avec $\xi_{ij} = x + \theta_{ij}\Delta x$, $0 < \theta_{ij} < 1$; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, r$. Si l'on fixe x et $x + \Delta x$, on aura :

$$\|F(x + \Delta x) - F(x)\| = \max_i \sum_{j=1}^{r} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \Delta x_k \right| \leqslant$$

$$\leqslant \max_i \sum_{j=1}^{r} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right| |\Delta x_k| \leqslant$$

$$\leqslant \max_k |\Delta x_k| \cdot \sum_{j=1}^{r} \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right| =$$

$$= r \|\Delta x\| \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right|.$$

Le nombre de couples (i, j) étant fini, il existe un couple (p, q) tel que

$$\max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{ij})}{\partial x_k} \right| = \sum_{k=1}^{n} \left| \frac{\partial f_{pq}(\xi_{pq})}{\partial x_k} \right| \leqslant \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial f_{ij}(\xi_{pq})}{\partial x_k} \right| = \|F'(\xi)\|,$$

où $\xi = \xi_{pq}$.

Ainsi

$$\|F(x + \Delta x) - F(x)\| \leqslant r \|\Delta x\| \, \|F'(\xi)\|,$$

ce qu'il fallait démontrer.

Corollaire 1. Si

$$f(x)=\begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix},$$

il vient

$$\| f(x+\Delta x)-f(x) \| \leqslant \| \Delta x \| \cdot \| f'(\xi) \|,$$

où $\xi = x+\theta\Delta x$ et $0<\theta<1$.
Ici $r=1$.

Corollaire 2. Avec $f(x) \in C^{(2)}$ on a:

$$\| f'(x+\Delta x)-f'(x) \| \leqslant n \| \Delta x \| \; \| f''(\xi) \|,$$

où $\xi = x+\theta\Delta x$ et $0<\theta<1$.

Lemme 2. *Si*

$$f(x)=\begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} \in C^{(2)}$$

dans un domaine convexe qui contient les points x et $x+\Delta x$, alors

$$\| f(x+\Delta x)-f(x)-f'(x)\Delta x \| \leqslant \frac{1}{2} n \| \Delta x \|^2 \cdot \| f''(\xi) \|, \tag{4}$$

où $\xi = x+\theta\Delta x$ *et* $0<\theta<1$.

Démonstration. En utilisant la formule du binôme de Taylor, on obtient:

$$\| f(x+\Delta x)-f(x)-f'(x)\Delta x \| =$$

$$= \| [f_i(x+\Delta x)-f_i(x)-df_i(x_i)] \| = \frac{1}{2} \left\| \left[\sum_{j,k} \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \Delta x_j \Delta x_k \right] \right\| \leqslant$$

$$\leqslant \frac{1}{2} \left\| \left[\sum_j |\Delta x_j| \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| |\Delta x_k| \right] \right\| \leqslant$$

$$\leqslant \frac{1}{2} \max_j |\Delta x_j| \cdot \max_k |\Delta x_k| \cdot \left\| \left[\sum_j \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \right] \right\| =$$

$$= \frac{1}{2} \| \Delta x \|^2 \left\| \left[\sum_j \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \right] \right\|, \tag{5}$$

où $\xi_i = x+\theta_i\Delta x, \quad 0<\theta_i<1$.

Puisque

$$\sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| \leqslant \max_{i,\,j} \sum_k \left| \frac{\partial^2 f_i(\xi_i)}{\partial x_j \partial x_k} \right| =$$

$$= \sum_k \left| \frac{\partial^2 f_p(\xi_p)}{\partial x_q \partial x_k} \right| \leqslant \max_{i,\,j} \sum_k \left| \frac{\partial^2 f_i(\xi_p)}{\partial x_j \partial x_k} \right| = \| f''(\xi_p) \|,$$

compte tenu du sens de la norme, l'inégalité (5) entraîne

$$\| f(x+\Delta x) - f(x) - f'(x)\,\Delta x \| \leqslant$$

$$\leqslant \frac{1}{2} \| \Delta x \|^2 \, [\| f''(\xi) \|] = \frac{n}{2} \| \Delta x \|^2 \, \| f''(\xi) \|,$$

où $\xi = \xi_p = x + \theta \Delta x$ et $0 < \theta < 1$.

§ 3*. Existence des solutions d'un système et convergence du processus de Newton

T h é o r è m e 1. *Soit un système réel d'équations algébriques ou transcendantes non linéaires*

$$f(x) = 0, \tag{1}$$

où le vecteur fonction

$$f(x) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \cdots\cdots\cdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix}$$

avec ses dérivées partielles premières et secondes est défini et continu dans un certain domaine ω, *c'est-à-dire*

$$f(x) \in C^{(2)}(\omega).$$

Posons que $x^{(0)}$ *est un point contenu dans* ω *avec son* $\mathscr{H}$*-voisinage fermé*

$$\overline{U}_{\mathscr{H}}(x^{(0)}) = \{\| x - x^{(0)} \| \leqslant \mathscr{H}\} \subset \omega,$$

*où par norme on entend la m-norme** (cf. chapitre VII, § 7) *et où l'on vérifie les conditions suivantes*:

* C'est-à-dire, si $A = [a_{ij}]$:

$$\| A \| = \| A \|_m = \max_i \sum_j | a_{ij} |.$$

1) *la matrice jacobienne* $W(x) = \left[\frac{\partial f_i}{\partial x_j}\right]$ *pour* $x = x^{(0)}$ *possède une inverse* $\Gamma_0 = W^{-1}(x^{(0)})$ *avec*

$$\| \Gamma_0 \| \leqslant A_0^* ;$$

2) $\| \Gamma_0 f(x^{(0)}) \| \leqslant B_0 \leqslant \frac{\mathscr{H}}{2}$;

3) $\sum_{k=1}^{n} \left| \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k} \right| \leqslant C$

pour $i, j = 1, 2, \ldots, n$ *et* $x \in \overline{U}_{\mathscr{H}}(x^{(0)})$;

4) *les constantes* A_0, B_0 *et* C *satisfont à l'inégalité*

$$\mu_0 = 2nA_0B_0C \leqslant 1. \tag{2}$$

Alors, pour une approximation initiale $x^{(0)}$, *le processus de Newton*

$$x^{(p+1)} = x^{(p)} - W^{-1}(x^{(p)}) f(x^{(p)}) \tag{3}$$

$(p = 0, 1, 2, \ldots)$ *converge et le vecteur limite*

$$x^* = \lim_{p \to \infty} x^{(p)}$$

est une solution du système (1) *telle que*

$$\| x^* - x^{(0)} \| \leqslant 2B_0 \leqslant \mathscr{H}.$$

D é m o n s t r a t i o n. Introduisons les notations

$$h_p = \| x^{(p+1)} - x^{(p)} \| = \max_k | x_k^{(p+1)} - x_k^{(p)} |,$$

$$\Gamma_p = W^{-1}(x^{(p)}) \quad (p = 0, 1, 2, \ldots).$$

La formule (3) entraîne

$$h_p = \| \Gamma_p f(x^{(p)}) \|.$$

Les conditions 1)-4) donnent les estimations des quantités Γ_p et $\Gamma_p f(x^{(p)})$.

Examinons d'abord le cas $p = 1$. En utilisant la condition 2), on a :

$$h_0 = \| x^{(1)} - x^{(0)} \| = \| W^{-1}(x^{(0)}) f(x^{(0)}) \| \leqslant B_0 \leqslant \frac{\mathscr{H}}{2} ;$$

---

* En d'autres termes, si $W(x^{(0)}) = [a_{ij}]$, alors $\Gamma_0 = W^{-1}(x^{(0)}) = \left[\frac{A_{ji}}{\Delta}\right]$, où A_{ij} sont les cofacteurs des éléments a_{ij} et $\Delta = \det [a_{ij}]$; par conséquent,

$$\| \Gamma_0 \| = \max_i \frac{1}{|\Delta|} \sum_{j=1}^{n} | A_{ji} |.$$

donc

$$h_0 \leqslant B_0$$

et

$$\overline{U}_{\frac{\mathcal{H}}{2}}(x^{(1)}) \subset \overline{U}_{\mathcal{H}}(x^{(0)}).$$

Pour évaluer $\Gamma_1 = W^{-1}(x^{(1)})$, appliquons la relation $(AB)^{-1} = B^{-1}A^{-1}$ pour mettre cette grandeur sous la forme

$$\Gamma_1 = [W(x^{(0)}) \cdot \Gamma_0 W(x^{(1)})]^{-1} = [\Gamma_0 W(x^{(1)})]^{-1} \cdot \Gamma_0. \tag{4}$$

En tenant compte de la condition 1) du théorème, on a:

$$\| E - \Gamma_0 W(x^{(1)}) \| = \| \Gamma_0 [W(x^{(0)}) - W(x^{(1)})] \| \leqslant$$
$$\leqslant \| \Gamma_0 \| \ \| W(x^{(0)}) - W(x^{(1)}) \| \leqslant A_0 \| W(x^{(1)}) - W(x^{(0)}) \|.$$

Puisque la condition (3) amène

$$\| f''(x) \| = \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial^2 f_i(x)}{\partial x_j \, \partial x_k} \right| \leqslant C,$$

en vertu du corollaire 2 du lemme 1 on obtient:

$$\| W(x^{(1)}) - W(x^{(0)}) \| = \| f'(x^{(1)}) - f'(x^{(0)}) \| \leqslant$$
$$\leqslant n \| x^{(1)} - x^{(0)} \| C \leqslant nB_0C;$$

et donc

$$\| E - \Gamma_0 W(x^{(1)}) \| \leqslant nA_0B_0C = \frac{\mu_0}{2} \leqslant \frac{1}{2}.$$

Par suite (chap. VII, § 10, théorème 5, corollaire), il existe une matrice inverse

$$[\Gamma_0 W(x^{(1)})]^{-1} = \{E - (E - \Gamma_0 W(x^{(1)}))\}^{-1},$$

et comme $\| E \| = \| E \|_m = 1$,

$$\| [\Gamma_0 W(x^{(1)})]^{-1} \| \leqslant \frac{1}{1 - \frac{\mu_0}{2}} \leqslant 2. \tag{5}$$

On déduit de la formule (4):

$$\| \Gamma_1 \| \leqslant \| [\Gamma_0 W(x^{(1)})]^{-1} \| \ \| \Gamma_0 \| \leqslant 2A_0 = A_1. \tag{6}$$

La formule (3) entraîne

$$f(x^{(0)} + f')(x^{(0)})(x^{(1)} - x^{(0)}) = 0,$$

d'où, en vertu du lemme 2,

$$\| f(x^{(1)}) \| = \| f(x^{(1)}) - f(x^{(0)}) - f'(x^{(0)})(x^{(1)} - x^{(0)}) \| \leqslant$$
$$\leqslant \frac{1}{2} n \| x^{(1)} - x^{(0)} \|^2 \ \| f''(\xi) \| \leqslant \frac{1}{2} nB_0^2C,$$

avec

$$\xi = x^{(0)} + \theta (x^{(1)} - x^{(0)}) \quad \text{et} \quad 0 < \theta < 1.$$

Compte tenu de l'inégalité (6), on obtient:

$$\| \Gamma_1 f(x^{(1)}) \| \leqslant \| \Gamma_1 \| \; \| f(x^{(1)}) \| \leqslant$$
$$\leqslant 2A_0 \cdot \frac{1}{2} n B_0^2 C = n A_0 B_0^2 C = \frac{1}{2} \mu_0 B_0 = B_1. \quad (7)$$

Ainsi pour le point $x^{(1)}$ nous avons

$$\overline{U}_{\frac{\mathcal{H}}{2}} (x^{(1)}) \subset \overline{U}_{\mathcal{H}} (x^{(0)}) \subset \omega$$

et, en outre,

$$\| \Gamma_1 \| \leqslant A_1, \qquad h_1 \| \Gamma_1 f(x^{(1)}) \| \leqslant B_1,$$

où

$$A_1 = 2A_0,$$
$$B_1 = \frac{1}{2} \mu_0 B_0 \leqslant \frac{\mathcal{H}}{4}.$$

Il en résulte

$$\mu_1 = 2nA_1 B_1 C = 2n \cdot 2A_0 \cdot \frac{1}{2} \mu_0 B_0 C = \mu_0 \cdot 2n A_0 B_0 C = \mu_0^2 \leqslant 1. \quad (8)$$

On retombe donc dans les conditions du théorème, à cette différence près qu'au lieu du voisinage $\overline{U}_{\mathcal{H}} (x^{(0)})$ on a le voisinage $\overline{U}_{\frac{\mathcal{H}}{2}} (x^{(1)})$ emboîté dans le premier voisinage.

En reprenant des raisonnements analogues, nous pouvons établir que les approximations successives $x^{(p)}$ $(p = 1, 2, \ldots)$ ont un sens et sont telles que

$$\overline{U}_{\mathcal{H}} (x^{(0)}) \supset \overline{U}_{\frac{\mathcal{H}}{2}} (x^{(1)}) \supset \ldots \supset \overline{U}_{\frac{\mathcal{H}}{2^p}} (x^{(p)}) \supset \ldots,$$

de plus

$$\| \Gamma_p \| = \| W^{-1} (x^{(p)}) \| \leqslant A_p,$$
$$\| \Gamma_p f(x^{(p)}) \| = \| x^{(p+1)} - x^{(p)} \| \leqslant B_p,$$

où les constantes A_p et B_p sont liées entre elles par les relations de récurrence

$$\left. \begin{aligned} A_p &= 2A_{p-1}, \\ B_p &= \frac{1}{2} \mu_{p-1} B_{p-1} \end{aligned} \right\} \quad (9)$$

et

$$\mu_p = 2nA_p B_p C \quad (p = 1, 2, \ldots). \quad (10)$$

Montrons que la suite des approximations $x^{(p)}$ $(p = 0, 1, 2, \ldots)$ vérifie le critère de Cauchy (chapitre VII, § 9). En effet, pour $q > 0$ on a:

$$x^{(p+q)} \in \overline{U}_{\frac{\mathscr{H}}{2^p}}(x^{(p)}).$$

Par suite, pour tout $\varepsilon > 0$ donné à l'avance

$$\| x^{(p+q)} - x^{(p)} \| \leqslant \frac{\mathscr{H}}{2^p} < \varepsilon,$$

si $p > N$ et $q > 0$ avec N suffisamment grand, ce qui est équivalent au critère de Cauchy. On en tire l'existence de la limite

$$\lim_{p \to \infty} x^{(p)} = x^* \in \overline{U}_{\mathscr{H}}(x^{(0)}).$$

Montrons maintenant que x^* est une solution du système (1). La relation (3) conduit à

$$f(x^{(p)}) + W(x^{(p)})(x^{(p+1)} - x^{(p)}) = 0.$$

En passant dans cette égalité à la limite quand $p \to \infty$ et en tenant compte du fait que

$$x^{(p+1)} - x^{(p)} \to 0,$$

ainsi que $W(x^{(p)})$ est continue et bornée dans $\overline{U}_{\mathscr{H}}(x^{(0)})$, on aura:

$$\lim_{p \to \infty} f(x^{(p)}) = 0.$$

On obtient en vertu de la continuité de $f(x)$:

$$f(\lim_{p \to \infty} x^{(p)}) = f(x^*) = 0,$$

c'est-à-dire x^* est une solution du système (1). En outre,

$$\| x^* - x^{(0)} \| = \left\| \sum_{p=0}^{\infty} [x^{(p+1)} - x^{(p)}] \right\| \leqslant$$

$$\leqslant \sum_{p=0}^{\infty} \| x^{(p+1)} - x^{(p)} \| \leqslant \sum_{p=0}^{\infty} B_p \leqslant B_0 + \frac{B_0}{2} + \ldots = 2B_0 \leqslant \mathscr{H}.$$

Le théorème est ainsi complètement démontré.

R e m a r q u e 1. Si $f(x) \in C^{(2)}(\omega)$ et dans le domaine ω le système (1) a une solution simple x^*, c'est-à-dire telle que

$$f(x^*) = 0, \quad f'(x^*) = W(x^*) \neq 0,$$

les conditions du théorème 1 seront évidemment respectées pour tout point $x^{(0)}$ suffisamment proche de x^*.

Pour vérifier la condition 2) il est utile de noter que B_0 donne une estimation de l'écart entre les approximations initiale et première du processus de Newton:

$$\| \Gamma_0 f(x^{(0)}) \| = \| x^{(1)} - x^{(0)} \| \leqslant B_0,$$

cette inégalité peut donc être vérifiée aisément dès qu'on trouve l'approximation $x^{(1)}$.

R e m a r q u e 2. On obtient des énoncés analogues du théorème de convergence si au lieu de la norme $\| A \|_m$ on recourt à la norme $\| A \|_l$ ou $\| A \|_k$.

§ 4*. Rapidité de la convergence d'un processus de Newton

T h é o r è m e 2. *Si les conditions* 1) *à* 4) *du théorème* 1 *du* § 3 *sont remplies, les approximations successives* $x^{(p)}$ ($p = 0, 1, 2, \ldots$) *vérifient l'inégalité*

$$\| x^* - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p - 1} B_0,$$

où x^* *est une solution du système et* μ_0 *est définie par la formule* (2) *du* § 3.

D é m o n s t r a t i o n. En appliquant les relations (9) et (10) du § 3, on a

$$\mu_p = 2nA_pB_pC = 2n \cdot 2A_{p-1} \cdot \frac{1}{2} \mu_{p-1} B_{p-1} \cdot C =$$
$$= \mu_{p-1} \cdot 2nA_{p-1}B_{p-1}C = \mu_{p-1}^2.$$

Il en résulte que

$$\left.\begin{array}{l} \mu_1 = \mu_0^2, \\ \mu_2 = \mu_1^2 = \mu_0^4, \\ \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \\ \mu_p = \mu_0^{2^p}. \end{array}\right\} \qquad (1)$$

Ensuite

$$B_p = \frac{1}{2} \mu_{p-1} B_{p-1} = \frac{1}{2} \mu_0^{2^{p-1}} B_{p-1}.$$

Donc

$$B_p = \frac{1}{2} \mu_0^{2^{p-1}} \cdot \frac{1}{2} \mu_0^{2^{p-2}} \ldots \frac{1}{2} \mu_0^{2^0} B_0 =$$
$$= \left(\frac{1}{2}\right)^p \cdot \mu_0^{2^{p-1} + 2^{p-2} + \ldots + 1} B_0 = \left(\frac{1}{2}\right)^p \mu_0^{2^p - 1} B_0. \qquad (2)$$

Comme

$$\| x^{(p+1)} - x^{(p)} \| \leqslant B_p,$$

on a pour $q > 1$

$$\| x^{(p+q)} - x^{(p)} \| \leqslant \| x^{(p+1)} - x^{(p)} \| + \| x^{(p+2)} - x^{(p+1)} \| + \ldots + \| x^{(p+q)} - x^{(p+q-1)} \| \leqslant$$
$$\leqslant B_p + B_{p+1} + \ldots + B_{p+q-1} =$$
$$= \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 + \left(\frac{1}{2}\right)^{p+1} \mu_0^{2^{p+1}-1} B_0 + \ldots +$$
$$+ \left(\frac{1}{2}\right)^{p+q-1} \mu_0^{2^{p+q-1}-1} B_0 = \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 \Big[1 +$$
$$+ \frac{1}{2} \cdot \mu_0^{2^p} + \ldots + \left(\frac{1}{2}\right)^{q-1} \mu_0^{2^p(2^{q-1}-1)} \Big].$$

On en déduit en tenant compte du fait que $\mu_0 \leqslant 1$

$$\| x^{(p+q)} - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} B_0 \Big[1 + \frac{1}{2} + \ldots + \left(\frac{1}{2}\right)^{q-1} \Big] \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0.$$

En passant à la limite quand $q \to \infty$, on obtient finalement:

$$\| x^* - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \mu_0^{2^p-1} B_0 \leqslant \left(\frac{1}{2}\right)^p \mu_0^{2^p-1} \mathscr{H},$$

où

$$\mu_0 = 2nA_0B_0C \leqslant 1.$$

Ainsi pour $\mu_0 < 1$ la convergence du processus de Newton est superrapide. En particulier, pour $p = 0$ on a:

$$\| x^* - x^{(0)} \| \leqslant 2B_0 \leqslant \mathscr{H}.$$

§ 5*. Unicité de la solution

Théorème 3. *Sous les conditions* 1) *à* 4) *du théorème* 1 *du* § 3, *le domaine*

$$\| x - x^{(0)} \| \leqslant 2B_0 \tag{1}$$

contient une seule solution du système (1) *du* § 3.

Démonstration. Supposons qu'en plus de la solution x^* du système (1) du § 3, définie par le processus de Newton, il existe une autre solution x^{**} de ce système telle que

$$\| x^{**} - x^{(0)} \| \leqslant 2B_0. \tag{2}$$

Les approximations successives $x^{(p)}$ ($p = 0, 1, 2, \ldots$) du processus de Newton sont comprises dans le voisinage de (1) et respectent la condition

$$f(x^{(p)}) + W_p (x^{(p+1)} - x^{(p)}) = 0$$

avec

$$W_p = W(x^{(p)}).$$

En tenant compte du fait que

$$f(x^{**}) = 0,$$

il vient

$$W_p(x^{(p+1)} - x^{**}) = f(x^{**}) - f(x^{(p)}) - W_p(x^{**} - x^{(p)})$$

et, par conséquent,

$$x^{(p+1)} - x^{**} = \Gamma_p[f(x^{**}) - f(x^{(p)}) - W_p(x^{**} - x^{(p)})],$$

où

$$\Gamma_p = W_p^{-1}.$$

Calculant l'estimation en norme, on aura:

$$\| x^{**} - x^{(p+1)} \| \leqslant \| \Gamma_p \| \, \| f(x^{**}) - f(x^{(p)}) - W_p(x^{**} - x^{(p)}) \|.$$

Dans les notations du § 3 (cf. théorème 1)

$$\| \Gamma_p \| \leqslant A_p.$$

L'application du lemme 2 du § 2 conduit à l'inégalité

$$\| f(x^{**}) - f(x^{(p)}) - W_p(x^{**} - x^{(p)}) \| \leqslant \frac{1}{2} nC \| x^{**} - x^{(p)} \|^2,$$

où la constante C est définie d'après la condition (3) du théorème 1. Par suite

$$\| x^{**} - x^{(p+1)} \| \leqslant \frac{1}{2} nA_pC \| x^{**} - x^{(p)} \|^2 \quad (p = 0, 1, 2, \ldots). \tag{3}$$

Posant dans l'inégalité (3) $p = 0$ et utilisant l'inégalité (2), on obtient

$$\| x^{**} - x^{(1)} \| \leqslant \frac{1}{2} nA_0C \| x^{**} - x^{(0)} \|^2 \leqslant 2nA_0B_0^2C,$$

ou, introduisant les nombres définis par les relations

$$\left.\begin{aligned} \mu_p &= 2nA_pB_pC, \\ B_{p+1} &= \frac{1}{2} \mu_p B_p \end{aligned}\right\} \quad (p = 0, 1, 2, \ldots), \tag{4}$$

on trouve

$$\| x^{**} - x^{(1)} \| \leqslant \mu_0 B_0 = 2B_1. \tag{5}$$

D'une façon analogue pour $p = 1$ on déduit des formules (3), (4) et (5):

$$\| x^{**} - x^{(2)} \| \leqslant \frac{1}{2} nA_1C \| x^{**} - x^{(1)} \|^2 \leqslant 2nA_1B_1^2C = \mu_1 B_1 = 2B_2.$$

En général,

$$\| x^{**} - x^{(p)} \| \leqslant 2B_p \quad (p = 0, 1, 2, \ldots). \tag{6}$$

Comme la formule (2) du § 4 entraîne que la grandeur $B_p \to 0$ quand $p \to \infty$, en passant à la limite dans l'inégalité (6), on a:

$$x^{**} = \lim_{p\to\infty} x^{(p)} = x^{*},$$

c'est-à-dire la solution du système (1) dans le domaine $\| x - x^{(0)} \| \leqslant \leqslant 2B_0$ est unique.

Remarque. Si le domaine $\overline{U}_{\mathcal{H}}(x^{(0)})$ est tel que

$$\frac{2}{\mu_0} B_0 \leqslant \mathcal{H},$$

le système (1) ne possède pas dans le domaine étendu (1)

$$\| x - x^{(0)} \| \leqslant \frac{2}{\mu_0} B_0 \tag{7}$$

d'autres solutions que x^{*}.

En effet, en supposant que le domaine (7) comporte une solution x^{**} du système (1) (§ 3) et en reprenant les raisonnements du théorème 1, on obtient une inégalité de la forme (3)

$$\| x^{**} - x^{(p+1)} \| \leqslant \frac{1}{2} nA_pC \| x^{**} - x^{(p)} \|^2,$$

où $x^{(p)}$ $(p = 0, 1, 2, \ldots)$ sont les approximations successives du processus de Newton à approximation initiale $x^{(0)}$. D'où, puisque

$$\| x^{**} - x^{(0)} \| \leqslant \frac{2}{\mu_0} B_0,$$

on a successivement, en utilisant les nombres $\mu_{p+1} = \mu_p^2$

$$\| x^{**} - x^{(1)} \| \leqslant \frac{1}{2} nA_0C \frac{4}{\mu_0^2} B_0^2 =$$

$$= 2nA_0B_0C \cdot \frac{1}{\mu_0^2} B_0 = \frac{1}{\mu_0} B_0 = \frac{2}{\mu_0^2} B_1 = \frac{2}{\mu_1} B_1,$$

$$\| x^{**} - x^{(2)} \| \leqslant \frac{1}{2} nA_1C \cdot \frac{4}{\mu_1^2} B_1^2 =$$

$$= 2nA_1B_1C \cdot \frac{1}{2} \mu_1 B_1 \cdot \frac{2}{\mu_1^3} = \mu_1 \cdot B_2 \cdot \frac{2}{\mu_1^3} = \frac{2}{\mu_1^2} B_2 = \frac{2}{\mu_2} B_2,$$

etc.

En général,

$$\| x^{**} - x^{(p)} \| \leqslant \frac{2}{\mu_p} B_p \quad (p = 0, 1, 2, \ldots).$$

Puisque

$$B_p = \frac{1}{2} \mu_{p-1} B_{p-1}$$

et

$$\mu_p = \mu_{p-1}^2,$$

il vient

$$\frac{B_p}{\mu_p} = \frac{1}{2} \cdot \frac{B_{p-1}}{\mu_{p-1}} = \left(\frac{1}{2}\right)^p \cdot \frac{B_0}{\mu_0}. \tag{8}$$

Cette dernière relation peut également s'obtenir directement des formules (1) et (2) du § 4.

Ainsi

$$\| x^{**} - x^{(p)} \| \leqslant \left(\frac{1}{2}\right)^{p-1} \frac{B_0}{\mu_0} \quad (p = 0, 1, 2, \ldots).$$

Par conséquent,

$$x^{**} = \lim_{p \to \infty} x^{(p)} = x^*,$$

ce qu'il fallait démontrer.

§ 6*. Stabilité de la convergence du processus de Newton devant la variation de l'approximation initiale

Théorème 4. *Si les conditions* 1)-4) *du théorème* 1 *du* § 3 *sont remplies et si*

$$\frac{2}{\mu_0} B_0 \leqslant \mathscr{H},$$

avec $\mu_0 = 2nA_0B_0C < 1$, *le processus de Newton converge vers la solution unique* x^* *du système* (1) (§ 3) *dans le domaine principal* $\| x - x^{(0)} \| \leqslant 2B_0$ *quel que soit le choix de l'approximation initiale* $x'^{(0)}$ *dans le domaine*

$$\| x'^{(0)} - x^{(0)} \| \leqslant \frac{1-\mu_0}{2\mu_0} B_0. \tag{1}$$

Démonstration. Par analogie avec les notations données ci-dessus

$$W_0 = W(x^{(0)}) \quad \text{et} \quad \Gamma_0 = W_0^{(-1)}$$

introduisons

$$W'_0 = W(x'^{(0)}) \quad \text{et} \quad \Gamma'_0 = (W'_0)^{-1}.$$

Montrons qu'au point $x'^{(0)}$ on vérifie des conditions analogues à 1)-4) du théorème 1.

Utilisant les notations et la méthode de démonstration du théorème 1, on a:

$$\| E - \Gamma_0 W'_0 \| = \| \Gamma_0 (W_0 - W'_0) \| \leqslant$$
$$\leqslant \| \Gamma_0 \| \, \| W_0 - W'_0 \| \leqslant A_0 nC \, \| x'^{(0)} - x^{(0)} \|.$$

D'où, compte tenu de l'inégalité (1),

$$\| E - \Gamma_0 W'_0 \| \leqslant A_0 nC \frac{1-\mu_0}{2\mu_0} B_0 = \frac{1-\mu_0}{4} \leqslant \frac{1}{4}.$$

Par suite,

$$\|(\Gamma_0 W_0')^{-1}\| = \|[E-(E-\Gamma_0 W_0')]^{-1}\| \leqslant$$
$$\leqslant \frac{1}{1-\|E-\Gamma_0 W_0'\|} \leqslant \frac{1}{1-\frac{1-\mu_0}{4}} = \frac{4}{3+\mu_0}. \quad (2)$$

Il existe donc

$$\Gamma_0' = (\Gamma_0 W_0')^{-1}\Gamma_0$$

et

$$\|\Gamma_0'\| \leqslant \|(\Gamma_0 W_0')^{-1}\| \, \|\Gamma_0\| \leqslant \frac{4A_0}{3+\mu_0} = A'. \quad (3)$$

Déduisons ensuite

$$\|\Gamma_0 f(x'^{(0)})\| \leqslant \|\Gamma_0\| \, \|f(x'^{(0)}) - f(x^{(0)}) -$$
$$- W_0(x'^{(0)} - x^{(0)})\| + \|\Gamma_0 f(x^{(0)})\| + \|x'^{(0)} - x^{(0)}\| \leqslant$$
$$\leqslant \frac{1}{2} A_0 nC \|x'^{(0)} - x^{(0)}\|^2 + B_0 + \|x'^{(0)} - x^{(0)}\| \leqslant$$
$$\leqslant \frac{1}{4}\mu_0 B_0 \frac{1-2\mu_0+\mu_0^2}{4\mu_0^2} + B_0 + \frac{1-\mu_0}{2\mu_0} B_0 =$$
$$= \frac{1-2\mu_0+\mu_0^2+16\mu_0+8-8\mu_0}{16\mu_0} B_0 = \frac{(3+\mu_0)^2}{16\mu_0} B_0.$$

On en tire en utilisant l'inégalité (2):

$$\|\Gamma_0' f(x'^{(0)})\| = \|(\Gamma_0 W_0)^{-1} \cdot \Gamma_0 f(x'^{(0)})\| \leqslant$$
$$\leqslant \|(\Gamma_0 W_0')^{-1}\| \cdot \|\Gamma_0 f(x'^{(0)})\| \leqslant$$
$$\leqslant \frac{4}{3+\mu_0} \cdot \frac{(3+\mu_0)^2}{16\mu_0} B_0 = \frac{3+\mu_0}{4\mu_0} B_0 = B'. \quad (4)$$

En vertu des inégalités (3) et (4), on obtient:

$$\mu' = 2nA'B'C = 2n \frac{4A_0}{3+\mu_0} \cdot \frac{3+\mu_0}{4\mu_0} B_0 C = 2nA_0 B_0 C \frac{1}{\mu_0} = 1.$$

De plus,

$$2B' + \|x'^{(0)} - x^{(0)}\| \leqslant \frac{3+\mu_0}{2\mu_0} B_0 + \frac{1-\mu_0}{2\mu_0} B_0 = \frac{2B_0}{\mu_0} \leqslant \mathscr{H}$$

et donc, à plus forte raison,

$$2B' \leqslant \frac{2B_0}{\mu_0} \leqslant \mathscr{H}.$$

Ainsi, au point $x'^{(0)}$ les conditions du théorème 1 sont complètement vérifiées; en outre

$$\bar{U}_{2B'}(x'^{(0)}) \subset \bar{U}_{\frac{2B_0}{\mu_0}}(x^{(0)}) \subset \bar{U}_{\mathscr{H}}(x^{(0)}) \quad (5)$$

(fig. 58).

La procédure de Newton

$$x'^{(p+1)} = x'^{(p)} - \Gamma'_p f(x'^{(p)}),$$

où

$$\Gamma'_p = W^{-1}(x'^{(p)}) \quad (p = 0, 1, 2, \ldots),$$

converge donc vers une certaine solution x'^* du système (1) du § 3 qui repose dans le domaine $\bar{U}_{2B'}(x'^{(0)})$. En vertu de la formule (5)

$$x'^* \in \bar{U}_{\frac{2B_0}{\mu\alpha_0^2}}(x^{(0)}).$$

Mais la remarque du théorème 3 du paragraphe précédent fait que dans le domaine $\bar{U}_{\frac{2B_0}{\mu_0}}(x^{(0)})$ il n'y a qu'une seule solution x^* du système principal (1). Donc

$$x'^* = x^*$$

et

$$x^* = \lim_{p\to\infty} x'^{(p)},$$

ce qu'il fallait démontrer.

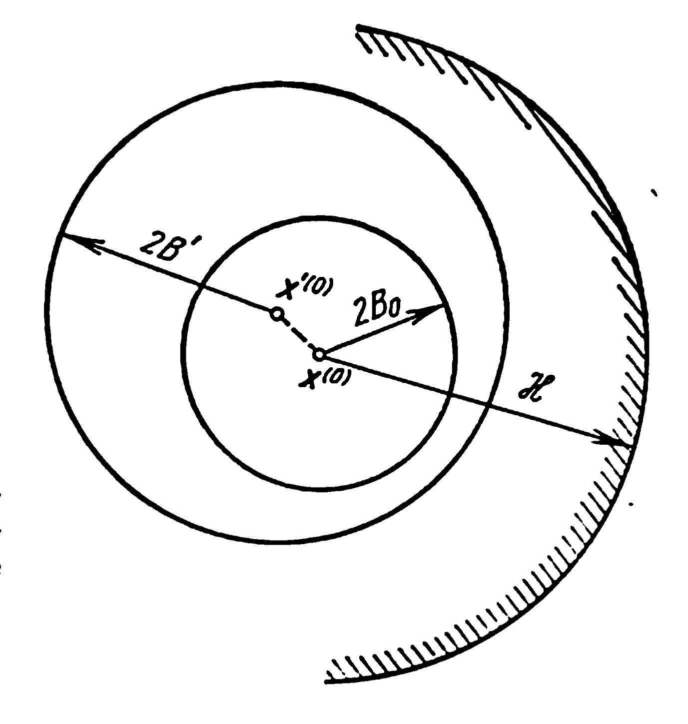


Fig. 58.

Remarque. Si $2B_0 < \mathscr{H}$ et $\mu_0 < 1$, pour la première approximation initiale $x^{(0)}$ il existe toujours un voisinage dont n'importe quel point peut être pris comme approximation initiale de la procédure de Newton qui converge vers la solution cherchée x^*.

En effet, soit

$$2B_0 < 2qB_0 = \mathscr{H},$$

où $q > 1$. Posant

$$\mu_0^* = \max\left(\mu_0, \frac{1}{q}\right),$$

on obtient en vertu des théorèmes 1 et 4 que pour une approximation initiale quelconque $x'^{(0)}$ qui vérifie la condition

$$\| x'^{(0)} - x^{(0)} \| \leqslant \frac{1-\mu_0^*}{2\mu_0} B_0$$

le processus de Newton correspondant converge vers la solution x^* du système (1).

§ 7. Méthode de Newton modifiée

La construction du processus de Newton

$$x^{(p+1)} = x^{(p)} - W^{-1}(x^{(p)}) f(x^{(p)}) \quad (p = 0, 1, 2, \ldots) \tag{1}$$

présente un inconvénient important qui consiste à calculer à chaque pas la matrice inverse $W^{-1}(x^{(p)})$. Si la matrice $W^{-1}(x)$ est continue dans le voisinage de la solution cherchée x^* et l'approximation initiale $x^{(0)}$ est suffisamment proche de x^*, on peut poser approximativement

$$W^{-1}(x^{(p)}) \approx W^{-1}(x^{(0)}),$$

et on retombe ainsi sur un *processus de Newton modifié*

$$\xi^{(p+1)} = \xi^{(p)} - W^{-1}(x^{(0)}) f(\xi^{(p)}) \tag{2}$$

$(p = 0, 1, 2, \ldots)$, où $\xi^{(0)} = x^{(0)}$. Remarquons que pour les processus (1) et (2) les premières approximations $x^{(1)}$ et $\xi^{(1)}$ coïncident

$$x^{(1)} = \xi^{(1)}.$$

La convergence du processus de Newton modifié (2) a été étudiée par L. Kantorovitch [1].

Théorème. *Si les conditions* 1) à 4) *du théorème* 1 (§ 3) *sont remplies et si*

$$\mu_0 = 2nA_0B_0C < 1,$$

le processus de Newton modifié (2) *déterminé par l'approximation initiale* $\xi^{(0)} = x^{(0)}$ *converge vers la solution* x^* *du système*

$$f(x) = 0$$

et

$$\| x^* - \xi^{(p)} \| \leqslant \mu_0^p \| x^* - x^{(0)} \| \leqslant 2B_0\mu_0^p \quad (p = 0, 1, 2, \ldots), \tag{3}$$

où on entend par norme la m-norme.

Démonstration. Considérons le vecteur fonction

$$F(x) = x - \Gamma_0 f(x) = [F_i(x)],$$

avec $\Gamma_0 = W^{-1}(x^{(0)})$.

Evidemment

$$F(\xi^{(p)}) = \xi^{(p)} - \Gamma_0 f(\xi^{(p)}) = \xi^{(p+1)} \quad (p = 0, 1, 2, \ldots). \tag{4}$$

De plus,

$$F'(x) = E - \Gamma_0 f'(x); \tag{5}$$

d'où, en particulier,

$$F'(x^{(0)}) = E - \Gamma_0 f'(x^{(0)}) = E - E = 0. \tag{6}$$

Montrons par récurrence que toute approximation $\xi^{(p)}$ ($p = 0, 1, 2, \ldots$) est comprise dans le voisinage $2B_0$ du point $x^{(0)}$

$$\| \xi^{(p)} - x^{(0)} \| < 2B_0. \qquad (7)$$

En effet, avec $p = 1$ l'égalité (7) est évidente du fait qu'en vertu de la condition (2) du théorème on a:

$$\| \xi^{(1)} - x^{(0)} \| = \| x^{(1)} - x^{(0)} \| \leqslant B_0.$$

Supposons maintenant que pour un certain p l'inégalité (7) soit vraie. Alors, en utilisant le lemme 2 du § 2 on a:

$$\| \xi^{(p+1)} - x^{(0)} \| = \| F(\xi^{(p)}) - x^{(0)} \| = \| \xi^{(p)} - \Gamma_0 f(\xi^{(p)}) - x^{(0)} \| =$$
$$= \| \Gamma_0 [f(\xi^{(p)}) - W(x^{(0)})(\xi^{(p)} - x^{(0)}] \| \leqslant \| \Gamma_0 f(x^{(0)}) \| +$$
$$+ \| \Gamma_0 \{ f(\xi^{(p)}) - f(x^{(0)}) - W(x^{(0)})(\xi^{(p)} - x^{(0)}) \} \| \leqslant$$
$$\leqslant B_0 + \frac{1}{2} A_0 n C \| \xi^{(p)} - x^{(0)} \|^2.$$

En appliquant l'inégalité (7) on trouve:

$$\| \xi^{(p+1)} - x^{(0)} \| < B_0 + \frac{1}{2} n A_0 C \cdot 4B_0^2 =$$
$$= B_0 + 2nA_0B_0C \cdot B_0 = (1 + \mu_0) B_0 < 2B_0,$$

ce qui démontre notre proposition.

Puisqu'on suppose que les conditions du théorème 1 du § 3 sont observées, le système $f(x) = 0$ possède une solution x^* telle que $\| x^* - x^{(0)} \| \leqslant 2B_0$.

Considérons la différence $x^* - \xi^{(p)}$, où $p \geqslant 1$. Compte tenu du fait que

$$F(x^*) \equiv x^* - \Gamma_0 f(x^*) = x^*$$

et en appliquant le lemme 1 du § 2, on a:

$$\| x^* - \xi^{(p)} \| = \| F(x^*) - F(\xi^{(p-1)}) \| \leqslant \| x^* - \xi^{(p-1)} \| \cdot \| F'(\theta) \|, \qquad (8)$$

où θ est un point du segment $[x^*, \xi^{(p-1)}]$.

Ensuite (cf. § 2, lemme 1, corollaire 2)

$$\| F'(\theta) \| = \| F'(\theta) - F'(x^{(0)}) \| \leqslant n \| \theta - x^{(0)} \| \max \| F''(\eta) \|, \qquad (9)$$

où η est un point du segment $[\theta, x^{(0)}]$. La formule (5) donne

$$F'(x) = \left[\delta_{ij} - \sum_{s=1}^{n} \gamma_{is} \frac{\partial f_s}{\partial x_j} \right],$$

où δ_{ij} est le symbole de Kronecker et $\Gamma_0 = [\gamma_{ij}]$. Donc

$$\frac{\partial F_i}{\partial x_j} = \delta_{ij} - \sum_{s=1}^{n} \gamma_{is} \frac{\partial f_s}{\partial x_j}$$

et

$$\frac{\partial^2 F_i}{\partial x_j \partial x_k} = -\sum_{s=1}^{n} \gamma_{is} \frac{\partial^2 f_s}{\partial x_j \partial x_k}.$$

Par suite,

$$\| F''(\eta) \| = \max_{i,j} \sum_{k=1}^{n} \left| \frac{\partial^2 F_i(\eta)}{\partial x_j \partial x_k} \right| = \max_{i,j} \sum_{k=1}^{n} \left| \sum_{s=1}^{n} \gamma_{is} \frac{\partial^2 f_s(\eta)}{\partial x_j \partial x_k} \right| \leqslant$$

$$\leqslant \max_{i,j} \sum_{s=1}^{n} |\gamma_{is}| \sum_{k=1}^{n} \left| \frac{\partial^2 f_s(\eta)}{\partial x_j \partial x_k} \right| \leqslant \max_{i,j} \sum_{s=1}^{n} |\gamma_{is}| C = C \| \Gamma_0 \| \leqslant A_0 C$$

et en vertu de (9), on a

$$\| F'(\theta) \| \leqslant n A_0 C \| \theta - x^{(0)} \|.$$

Le point θ appartient évidemment au voisinage $2B_0$ du point $x^{(0)}$; donc

$$\| \theta - x^{(0)} \| \leqslant 2B_0$$

et

$$\| F'(\theta) \| \leqslant 2nA_0B_0C = \mu_0. \qquad (10)$$

Si l'on tient compte de l'inégalité (10), l'inégalité (8) permet de déduire

$$\| x^* - \xi^{(p)} \| \leqslant \mu_0 \| x^* - \xi^{(p-1)} \|,$$

d'où

$$\| x^* - \xi^{(p)} \| \leqslant \mu_0^p \| x^* - \xi^{(0)} \| = \mu_0^p \| x^* - x^{(0)} \| \leqslant 2B\mu_0^p.$$

Pour $\mu_0 < 1$, la dernière inégalité entraîne

$$\lim_{p \to \infty} \xi^{(p)} = x^*.$$

Le théorème est complètement démontré.

§ 8. Méthode des approximations successives

Soit un système d'équations non linéaires de forme

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots\ldots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \qquad (1)$$

où les fonctions $\varphi_1, \varphi_2, \ldots, \varphi_n$ sont réelles, définies et continues dans un certain voisinage ω d'une solution isolée $(x_1^*, x_2^*, \ldots, x_n^*)$ de ce système.

Introduisons dans la discussion les vecteurs

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \quad \text{et} \quad \boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_n(\boldsymbol{x})),$$

on peut alors écrire le système (1) sous une forme plus compacte

$$\boldsymbol{x} = \boldsymbol{\varphi}(\boldsymbol{x}). \tag{2}$$

Pour trouver le vecteur racine $\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ de l'équation (2), il est souvent commode d'utiliser la *méthode des approximations successives*

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{\varphi}(\boldsymbol{x}^{(p)}) \qquad (p = 0, 1, 2, \ldots), \tag{3}$$

où l'approximation initiale $\boldsymbol{x}^{(0)} \approx \boldsymbol{x}^*$. La convergence de ce processus sera étudiée dans ce qui suit. Constatons que si le processus itératif (3) converge, la valeur limite

$$\boldsymbol{\xi} = \lim_{p \to \infty} \boldsymbol{x}^{(p)} \tag{4}$$

sera nécessairement une racine de l'équation (2). En effet, en supposant la relation (4) respectée et en passant à la limite dans l'égalité (3) quand $p \to \infty$, on a en vertu de la continuité de la fonction $\boldsymbol{\varphi}(\boldsymbol{x})$

$$\lim_{p \to \infty} \boldsymbol{x}^{(p+1)} = \boldsymbol{\varphi}(\lim_{p \to \infty} \boldsymbol{x}^{(p)}),$$

c'est-à-dire

$$\boldsymbol{\xi} = \boldsymbol{\varphi}(\boldsymbol{\xi}).$$

Ainsi $\boldsymbol{\xi}$ est une racine de l'équation vectorielle (2).

Si, en outre, toutes les approximations $\boldsymbol{x}^{(p)}$ $(p = 0, 1, 2, \ldots)$ appartiennent au domaine ω et si $\boldsymbol{x}^*$ est une solution u n i q u e du système (2) dans ω, alors, évidemment,

$$\boldsymbol{\xi} = \boldsymbol{x}^*.$$

La méthode des approximations successives peut être appliquée également au système général

$$\boldsymbol{f}(\boldsymbol{x}) = 0, \tag{5}$$

où $\boldsymbol{f}(\boldsymbol{x})$ est un vecteur fonction défini et continu dans le voisinage ω du vecteur solution isolé $\boldsymbol{x}^*$. Par exemple, récrivons ce système sous la forme suivante:

$$\boldsymbol{x} = \boldsymbol{x} + \Lambda \boldsymbol{f}(\boldsymbol{x}),$$

avec Λ une matrice régulière. Introduisant les notations

$$\boldsymbol{x} + \Lambda \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\varphi}(\boldsymbol{x}), \tag{6}$$

on aura

$$\boldsymbol{x} = \boldsymbol{\varphi}(\boldsymbol{x}). \tag{7}$$

A cette dernière équation il est facile d'appliquer la méthode des approximations successives ordinaire (3).

Si la fonction $f(x)$ possède une dérivée continue $f'(x)$ dans ω, la formule (6) entraîne

$$\varphi'(x) = E + \Lambda f'(x).$$

Dans les paragraphes qui suivent nous montrerons que pour l'équation (7) le processus itératif converge rapidement si $\varphi'(x)$ est petit en norme. Compte tenu de cette circonstance, choisissons la matrice Λ telle que

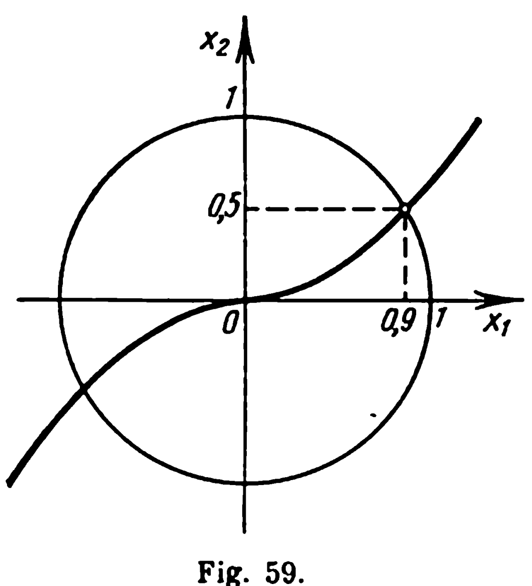


Fig. 59.

$$\varphi'(x^{(0)}) = E + \Lambda f'(x^{(0)}) = 0;$$

d'où, si la matrice $f'(x^{(0)})$ est régulière,

$$\Lambda = -[f'(x^{(0)})]^{-1}.$$

Remarquons qu'au fond c'est un processus de Newton modifié appliqué à l'équation (5) (cf. § 7).

Dans le cas du det $f'(x^{(0)}) = 0$, il convient de choisir une autre approximation initiale $x^{(0)}$.

Il existe d'autres modes encore pour remplacer le système (5) par un système (7) équivalent.

Exemple. Résoudre à l'aide de la méthode des approximations successives la solution approchée du système

$$\left.\begin{aligned} x_1^2 + x_2^2 &= 1, \\ x_1^3 - x_2 &= 0. \end{aligned}\right\} \qquad (8)$$

Solution. La courbe de la figure 59 montre que le système (8) possède deux solutions qui ne diffèrent que par le signe. Bornons-nous à rechercher la solution positive. D'après le dessin nous pouvons prendre pour approximation initiale de (8):

$$x^{(0)} = \begin{bmatrix} 0,9 \\ 0,5 \end{bmatrix}.$$

Posant

$$f(x) = \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ x_1^3 - x_2 \end{bmatrix},$$

on aura:

$$f'(x) = \begin{bmatrix} 2x_1 & 2x_2 \\ 3x_1^2 & -1 \end{bmatrix}.$$

D'où

$$f'(x^{(0)}) = \begin{bmatrix} 1{,}8 & 1 \\ 2{,}43 & -1 \end{bmatrix}$$

et

$$\det f'(x^{(0)}) = -1{,}8 - 2{,}43 = -4{,}23.$$

La matrice $f'(x^{(0)})$ étant régulière, il existe une inverse

$$[f'(x^{(0)})]^{-1} = -\frac{1}{4{,}23} \begin{bmatrix} -1 & -1 \\ -2{,}43 & 1{,}8 \end{bmatrix}.$$

Ainsi

$$\Lambda = -[f'(x^{(0)})]^{-1} = \frac{1}{4{,}23} \begin{bmatrix} -1 & -1 \\ -2{,}43 & 1{,}8 \end{bmatrix}.$$

Posons

$$\varphi(x) = x + \Lambda f(x) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{4{,}23} \begin{bmatrix} 1 & 1 \\ 2{,}43 & -1{,}8 \end{bmatrix} \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ x_1^3 - x_2 \end{bmatrix}.$$

Le système (8) sera alors équivalent à l'équation matricielle normalisée

$$x = \varphi(x). \tag{9}$$

Utilisons la formule (4) pour trouver les approximations successives de la solution du système (9):

$$x^{(1)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} - \frac{1}{4{,}23} \begin{bmatrix} 1 & 1 \\ 2{,}43 & 1{,}8 \end{bmatrix} \begin{bmatrix} x_1^{(0)2} + x_2^{(0)2} - 1 \\ x_1^{(0)3} - x_2^{(0)} \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}9 \\ 0{,}5 \end{bmatrix} - \frac{1}{4{,}23} \begin{bmatrix} 1 & 1 \\ 2{,}43 & -1{,}8 \end{bmatrix} \begin{bmatrix} 0{,}060 \\ 0{,}229 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}9 \\ 0{,}5 \end{bmatrix} - \begin{bmatrix} 0{,}0683 \\ -0{,}0630 \end{bmatrix} = \begin{bmatrix} 0{,}8317 \\ 0{,}5630 \end{bmatrix};$$

$$x^{(2)} = \begin{bmatrix} 0{,}8317 \\ 0{,}5630 \end{bmatrix} - \frac{1}{4{,}23} \begin{bmatrix} 1 & 1 \\ 2{,}43 & -1{,}8 \end{bmatrix} \begin{bmatrix} 0{,}8317^2 + 0{,}5630^2 - 1 \\ 0{,}8317^3 - 0{,}5630 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}8317 \\ 0{,}5630 \end{bmatrix} - \begin{bmatrix} -0{,}0049 \\ 0{,}0003 \end{bmatrix} = \begin{bmatrix} 0{,}8268 \\ 0{,}5633 \end{bmatrix};$$

$$x^{(3)} = \begin{bmatrix} 0{,}8268 \\ 0{,}5633 \end{bmatrix} - \begin{bmatrix} 0{,}0007 \\ 0{,}0002 \end{bmatrix} = \begin{bmatrix} 0{,}8261 \\ 0{,}5631 \end{bmatrix};$$

$$x^{(4)} = \begin{bmatrix} 0{,}8261 \\ 0{,}5631 \end{bmatrix} - \begin{bmatrix} 0{,}0000 \\ -0{,}0005 \end{bmatrix} = \begin{bmatrix} 0{,}8261 \\ 0{,}5636 \end{bmatrix},$$

etc.

En s'arrêtant à la quatrième approximation, on obtient la solution

$$x_1 = 0{,}8261\,; \quad x_2 = 0{,}5636,$$

et

$$f(x) = \begin{bmatrix} 0{,}0000 \\ 0{,}0002 \end{bmatrix}$$

à 10^{-4} près.

§ 9*. Notion de l'application contractante

Soit un système non linéaire

$$\left.\begin{aligned} y_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ y_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots \\ y_n &= \varphi_n(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \tag{1}$$

où les fonctions $\varphi_1, \varphi_2, \ldots, \varphi_n$ sont définies et continues dans un domaine connu G d'un espace réel E_n à n dimensions (chapitre X, § 1); les valeurs de ces fonctions $(y_1, y_2, \ldots, y_n)$ pour $(x_1, x_2, \ldots, x_n) \in G$ remplissent un certain domaine $G' \subset E_n$. Le système (1) établit une *application* du domaine G sur le domaine G'. Introduisons les vecteurs

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \qquad \varphi = \begin{bmatrix} \varphi_1 \\ \cdot \\ \cdot \\ \cdot \\ \varphi_n \end{bmatrix}$$

pour mettre le système (1) sous la forme abrégée:

$$y = \varphi(x). \tag{1'}$$

Introduisons dans l'espace E_n une norme canonique $\|x\|$ vérifiant les conditions ordinaires. On peut poser, par exemple,

$$\|x\|_m = \max_i |x_i|,$$

soit

$$\|x\|_l = \sum_i |x_i|,$$

soit encore

$$\|x\|_k = \sqrt{\sum x_i^2},$$

L'application (1) ou (1′) s'appelle *contractante* dans le domaine G s'il existe un nombre $q < 1$ tel que pour les deux points quelconques $x_1, x_2 \in G$, leurs transformés $y_1 = \varphi(x_1)$ et $y_2 = \varphi(x_2)$ satisfassent à la condition

$$\| y_1 - y_2 \| \leqslant q \| x_1 - x_2 \|, \tag{2}$$

c'est-à-dire

$$\| \varphi(x_1) - \varphi(x_2) \| \leqslant q \| x_1 - x_2 \| \quad (0 \leqslant q < 1). \tag{2'}$$

Considérons l'équation vectorielle non linéaire

$$x = \varphi(x), \tag{3}$$

équivalente au système d'équations non linéaire de forme spéciale

$$\left.\begin{array}{l} x_1 = \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 = \varphi_2(x_1, x_2, \ldots, x_n), \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ x_n = \varphi_n(x_1, x_2, \ldots, x_n). \end{array}\right\} \tag{3'}$$

La solution x^* de cette équation, si elle existe, est un *point fixe* de la transformation (1). Pour trouver x^* construisons le *processus itératif*

$$x^{(p)} = \varphi(x^{(p-1)}) \quad (p = 1, 2, \ldots), \tag{4}$$

où $x^{(0)} \in G$.

Théorème 1. *Supposons que le domaine G soit fermé et que l'application* (1) *soit contractante dans G, c'est-à-dire que la condition* (2) *soit respectée. Si dans ce cas pour le processus itératif* (4) *toutes les approximations successives $x^{(p)} \in G$ ($p = 0, 1, 2, \ldots$),* 1) *le processus* (4) *est convergent indépendamment du choix de l'approximation initiale $x^{(0)}$, c'est-à-dire il existe une solution*

$$x^* = \lim_{p \to \infty} x^{(p)}; \tag{5}$$

2) *le vecteur limite x^* est solution unique de l'équation* (3) *dans le domaine G;* 3) *l'estimation*

$$\| x^* - x^{(p)} \| \leqslant \frac{q^p}{1-q} \| x^{(1)} - x^{(0)} \| \tag{6}$$

est vraie.

Démonstration. 1) Pour prouver la convergence de la suite des approximations $x^{(p)}$ ($p = 0, 1, 2, \ldots$), appliquons le *critère de Cauchy* (cf. chapitre VII, § 9). On a

$$\begin{aligned} \| x^{(p+k)} - x^{(p)} \| &= \| (x^{(p+1)} - x^{(p)}) + (x^{(p+2)} - x^{(p+1)}) + \ldots + \\ &\quad + (x^{(p+k)} - x^{(p+k-1)} \| \leqslant \| x^{(p+1)} - x^{(p)} \| + \\ &\quad + \| x^{(p+2)} - x^{(p+1)} \| + \ldots + \| x^{(p+k)} - x^{(p+k-1)} \|. \end{aligned} \tag{7}$$

Utilisons la relation (4) et la «condition de contraction» (2') pour obtenir successivement

$$\|x^{(s+1)}-x^{(s)}\|=\|\varphi(x^{(s)})-\varphi(x^{(s-1)})\|\leqslant q\|x^{(s)}-x^{(s-1)}\|\leqslant$$
$$\leqslant q^2\|x^{(s-1)}-x^{(s-2)}\|\leqslant q^s\|x^{(1)}-x^{(0)}\| \quad (8)$$

avec $s\geqslant 0$. C'est pourquoi en renforçant le deuxième membre de l'inégalité (7), on aura

$$\|x^{(p+k)}-x^{(p)}\|\leqslant q^p\|x^{(1)}-x^{(0)}\|+q^{p+1}\|x^{(1)}-x^{(0)}\|+$$
$$+\ldots+q^{p+k-1}\|x^{(1)}-x^{(0)}\|$$

ou, en recourant à la formule de la somme des termes d'une progression géométrique,

$$\|x^{(p+k)}-x^{(p)}\|\leqslant\frac{q^p-q^{p+k}}{1-q}\|x^{(1)}-x^{(0)}\|\leqslant\frac{q^p}{1-q}\|x^{(1)}-x^{(0)}\|. \quad (9)$$

Comme $0\leqslant q<1$ et, par conséquent, $q^p\to 0$ lorsque $p\to\infty$ il apparaît de la formule (9) que pour tout $\varepsilon>0$ il existe $N=N(\varepsilon)$ telle qu'avec $p>N(\varepsilon)$ et $k>0$ l'inégalité

$$\|x^{(p+k)}-x^{(p)}\|<\varepsilon$$

est vraie, c'est-à-dire que le critère de Cauchy de la suite $x^{(p)}$ $(p=0, 1, 2, \ldots)$ est observé. Il existe donc une limite

$$x^*=\lim_{p\to\infty}x^{(p)},$$

et $x^*\in G$ du fait que le domaine G est fermé.

2) Le vecteur x^* est une solution de l'équation (3) du fait qu'en passant à la limite quand $p\to\infty$ dans l'égalité (4) et compte tenu de la continuité dans G du vecteur fonction $\varphi(x)$, on aura

$$\lim_{p\to\infty}x^{(p)}=\varphi(\lim_{p\to\infty}x^{(p-1)}),$$

c'est-à-dire

$$x^*=\varphi(x^*). \quad (10)$$

Cette solution est unique dans G. En effet, supposons que $x^{*\prime}$ soit une autre solution de l'équation (3)

$$x^{*\prime}=\varphi(x^{*\prime}). \quad (11)$$

En retranchant l'égalité (11) de l'égalité (10), on obtient:

$$x^*-x^{*\prime}=\varphi(x^*)-\varphi(x^{*\prime})$$

pour en tirer

$$\|x^*-x^{*\prime}\|=\|\varphi(x^*)-\varphi(x^{*\prime})\|\leqslant q\|x^*-x^{*\prime}\|$$

ou

$$(1-q)\|x^*-x^{*\prime}\|\leqslant 0. \quad (12)$$

Comme $1 - q > 0$, l'inégalité (12) ne peut avoir lieu que pour $\| x^* - x^{*\prime} \| = 0$, c'est-à-dire si $x^* = x^{*\prime}$. Ainsi, dans le domaine G l'équation (3) ne peut avoir d'autre solution.

3) En passant à la limite dans l'inégalité (9) quand $k \to \infty$, on obtient l'estimation (6).

Le théorème 1 est complètement démontré.

R e m a r q u e 1. Si le domaine G coïncide avec l'espace E_n tout entier, la condition $x^{(p)} \in G$ $(p = 0, 1, 2, \ldots)$ devient évidemment inutile.

R e m a r q u e 2. Si l'on utilise les inégalités

$$\| x^{(p+1)} - x^{(p)} \| \leqslant q \| x^{(p)} - x^{(p-1)} \|,$$
$$\| x^{(p+2)} - x^{(p+1)} \| \leqslant q^2 \| x^{(p)} - x^{(p-1)} \|,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

la formule (7) donne

$$\| x^{(p+k)} - x^{(p)} \| \leqslant q \| x^{(p)} - x^{(p-1)} \| + q^2 \| x^{(p)} - x^{(p-1)} \| + \\ + \ldots + q^k \| x^{(p)} - x^{(p-1)} \| \leqslant \frac{q}{1-q} \| x^{(p)} - x^{(p-1)} \|.$$

D'où quand $k \to \infty$

$$\| x^* - x^{(p)} \| \leqslant \frac{q}{1-q} \| x^{(p)} - x^{(p-1)} \|. \tag{13}$$

En particulier, si $0 \leqslant q \leqslant \frac{1}{2}$, la formule (13) entraîne que pour

$$\| x^{(p)} - x^{(p-1)} \| \leqslant \varepsilon$$

l'inégalité

$$\| x^* - x^{(p)} \| \leqslant \varepsilon$$

est vérifiée.

Les conditions du théorème 1 imposent que t o u t e s les approximations $x^{(p)}$ appartiennent au domaine fixé G. Dans la pratique cette condition est parfois difficile à vérifier. Aussi donnerons-nous un théorème légèrement modifié.

T h é o r è m e 2. *Supposons que l'application* (1) *soit contractante dans le domaine fermé G et que g soit un domaine borné compris dans G avec son voisinage* ρ (au sens de la norme adoptée), *où*

$$\rho \geqslant \frac{Dq}{1-q}, \tag{14}$$

D étant le diamètre du domaine g et q le coefficient correspondant de l'inégalité (2). *Alors, si*

$$x^{(0)} \in g \quad \text{et} \quad x^{(1)} = \varphi(x^{(0)}) \in g,$$

le processus itératif (4) *converge et les conclusions du théorème* 1 *sont valables.*

Démonstration. Il est clair qu'il suffit de montrer que toutes les approximations successives sont comprises dans le domaine G:

$$x^{(p)} \in G \qquad (p=2,\ 3,\ \ldots). \tag{15}$$

Donnons la démonstration par récurrence.

Par hypothèse $x^{(0)} \in G$ et $x^{(1)} \in G$. Soit maintenant $x^{(s)} \in G$ $(s=0,\ 1,\ \ldots,\ p)$. Comme

$$\| x^{(1)} - x^{(0)} \| \leqslant D,$$

en vertu de l'inégalité (8), on a:

$$\| x^{(s+1)} - x^{(s)} \| \leqslant q^s \| x^{(1)} - x^{(0)} \| \leqslant Dq^s \quad (s=0,\ 1,\ 2.\ \ldots,\ p).$$

Par conséquent,

$$\| x^{(p+1)} - x^{(1)} \| \leqslant \| x^{(2)} - x^{(1)} \| + \ldots + \| x^{(p+1)} - x^{(p)} \| \leqslant$$
$$\leqslant Dq + \ldots + Dq^p \leqslant \frac{Dq}{1-q} \leqslant \rho.$$

Donc $x^{(p+1)} \in G$.

Ainsi la propriété (15) est établie et nous tombons dans les conditions du théorème (1). Le théorème (2) est démontré.

Remarque. Les conditions du théorème 2 sont équivalentes aux conditions suivantes: soit $x^{(0)} \in G$ et $x^{(1)} \in G$; en outre, $\| x^{(1)} - x^{(0)} \| \leqslant D$. Si la distance du point $x^{(1)}$ à la frontière Γ du domaine G est égale ou inférieure à $\frac{Dq}{1-q}$, le processus itératif (4) converge.

§ 10*. Première condition suffisante de convergence des approximations successives

Considérons le système réduit réel

$$\left.\begin{aligned} x_1 &= \varphi_1(x_1,\ x_2,\ \ldots,\ x_n), \\ x_2 &= \varphi_2(x_1,\ x_2,\ \ldots,\ x_n), \\ &\ldots\ldots\ldots\ldots \\ x_n &= \varphi_n(x_1,\ x_2,\ \ldots,\ x_n), \end{aligned}\right\} \tag{1}$$

ou sous forme vectorielle

$$x = \varphi(x), \tag{1'}$$

avec

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \vdots \\ \varphi_n(x) \end{bmatrix}.$$

On suppose que le vecteur fonction $\varphi(x)$ est défini et continu avec sa dérivée $\varphi'(x) = \left[\frac{\partial \varphi_i}{\partial x_j}\right]$ dans un domaine fermé borné convexe $G \subset E_n$.

Dans ce paragraphe nous utiliserons deux normes:

$$\| x \|_m = \max_i | x_i |$$

et

$$\| x \|_l = \sum_{i=1}^{n} | x_i |.$$

Introduisons les normes par rapport au domaine G:

$$\| \varphi'(x) \|_{\mathrm{I}} = \max_{x \in G} \| \varphi'(x) \|_m \tag{2}$$

et

$$\| \varphi'(x) \|_{\mathrm{II}} = \max_{x \in G} \| \varphi'(x) \|_l, \tag{3}$$

où

$$\| \varphi'(x) \|_m = \max_i \sum_{j=1}^{n} \left| \frac{\partial \varphi_i(x)}{\partial x_j} \right| \tag{2'}$$

et

$$\| \varphi'(x) \|_l = \max_j \sum_{i=1}^{n} \left| \frac{\partial \varphi_i(x)}{\partial x_j} \right|. \tag{3'}$$

Théorème. Soient les fonctions $\varphi(x)$ et $\varphi'(x)$ *continues dans le domaine G, l'inégalité*

$$\| \varphi'(x) \|_{\mathrm{I}} \leqslant q < 1, \tag{4}$$

où q est une certaine constante, étant vérifiée dans G.

Si les approximations successives

$$x^{(p+1)} = \varphi(x^{(p)}) \tag{5}$$

($p = 0, 1, 2, \ldots$) *ne sortent pas hors du domaine G, le processus itératif* (5) *converge et dans le domaine G le vecteur limite*

$$x^* = \lim_{p \to \infty} x^{(p)}$$

est la solution unique du système (1).

Démonstration. En vertu du théorème 1 du paragraphe précédent, il suffit de montrer que sous la condition (4) l'application

$$y = \varphi(x) \tag{6}$$

est contractante dans le domaine G au sens de la m-norme.

Soit $x_1, x_2 \in G$ et $y_i = \varphi(x_i)$ $(i = 1, 2)$. Le corollaire 1 du lemme 1 du § 2 entraîne

$$\| y_1 - y_2 \|_m = \| \varphi(x_1) - \varphi(x_2) \|_m \leqslant$$
$$\leqslant \| x_1 - x_2 \|_m \| \varphi'(\xi) \|_m \leqslant \| x_1 - x_2 \|_m \| \varphi'(x) \|_{\mathrm{I}}.$$

On en tire

$$\| y_1 - y_2 \|_m \leqslant q \| x_1 - x_2 \|_m,$$

où $0 \leqslant q < 1$, ce qu'il fallait démontrer.

Corollaire. Le processus itératif (5) converge si

$$\sum_{j=1}^{n} \left| \frac{\partial \varphi_i(x)}{\partial x_j} \right| \leqslant q_i < 1 \quad (i = 1, 2, \ldots, n) \tag{7}$$

pour $x \in G$.

Il est évident que le système des inégalités (7) entraîne la condition (4) du théorème.

Remarque. Le théorème 1 du § 9 conduit pour l'approximation $x^{(p)}$ à l'estimation suivante

$$\| x^* - x^{(p)} \|_m \leqslant \frac{q^p}{1-q} \| x^{(1)} - x^{(0)} \|_m \quad (p = 0, 1, 2, \ldots),$$

où $x^{(1)} = \varphi(x^{(0)})$.

§ 11*. Deuxième condition suffisante de convergence des approximations successives

Avant de donner la démonstration du théorème de convergence qui fait appel aux normes $\| \varphi'(x) \|_{\mathrm{II}}$, nous déduirons une estimation de la différence des valeurs du vecteur fonction analogue au théorème de la moyenne et qui présente elle-même un intérêt propre.

Lemme. *Si le vecteur fonction*

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

est continu avec sa dérivée $f'(x)$ dans un domaine convexe qui contient les points x et $x + \Delta x$, alors

$$\| f(x + \Delta x) - f(x) \|_l \leqslant \| \Delta x \|_l \cdot \| f'(\xi) \|_l, \tag{1}$$

où $\xi = x + \theta \Delta x$ et $0 < \theta < 1$.

Démonstration. Considérons la fonction auxiliaire

$$\Phi(t) = \sum_{i=1}^{n} \varepsilon_i [f_i(x + t\Delta x) - f_i(x)],$$

où $0 \leqslant t \leqslant 1$ est argument scalaire et ε_i un système de nombres qui prennent les valeurs -1, 0, 1. Evidemment, $\Phi(0) = 0$. Appliquant le théorème de Lagrange des accroissements finis, on obtient :

$$\sum_{i=1}^{n} \varepsilon_i [f_i(x+\Delta x)-f_i(x)] = \Phi(1)-\Phi(0) = \Phi'(\theta) =$$

$$= \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{n} \frac{\partial f_i(\xi)}{\partial x_j} \Delta x_j,$$

avec $\xi = x+\theta\Delta x$ et $0<\theta<1$.

On en tire, compte tenu du fait que $|\varepsilon_i| \leqslant 1$:

$$\sum_{i=1}^{n} \varepsilon_i [f_i(x+\Delta x)-f_i(x)] \leqslant$$

$$\leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} \left|\frac{\partial f_i(\xi)}{\partial x_j}\right| \cdot |\Delta x_j| = \sum_{j=1}^{n} |\Delta x_j| \sum_{i=1}^{n} \left|\frac{\partial f_i(\xi)}{\partial x_j}\right|. \qquad (2)$$

Puisque

$$\sum_{i=1}^{n} \left|\frac{\partial f_i(\xi)}{\partial x_j}\right| \leqslant \max_j \sum_{i=1}^{n} \left|\frac{\partial f_i(\xi)}{\partial x_j}\right| = \|f'(\xi)\|_l,$$

en renforçant l'inégalité (2), on obtient :

$$\sum_{i=1}^{n} \varepsilon_i [f_i(x+\Delta x)-f_i(x)] \leqslant \sum_{j=1}^{n} |\Delta x_j| \, \|f'(\xi)\|_l =$$

$$= \|f'(\xi)\|_l \cdot \sum_{j=1}^{n} |\Delta x_j| = \|f'(\xi)\|_l \cdot \|\Delta x\|_l.$$

Posons dans cette dernière inégalité

$$\varepsilon_i = \operatorname{sgn} [f_i(x+\Delta x)-f_i(x)] \quad (i=1, 2, \ldots, n)$$

pour trouver finalement :

$$\sum_{i=1}^{n} |f_i(x+\Delta x)-f_i(x)| \leqslant \|f'(\xi)\|_l \, \|\Delta x\|_l,$$

c'est-à-dire

$$\|f(x+\Delta x)-f(x)\|_l \leqslant \|\Delta x\|_l \, \|f'(\xi)\|_l, \qquad (2')$$

ce qu'il fallait démontrer *.

* Si l'on applique directement le théorème de la moyenne à chaque composante du vecteur $f(x+\Delta x) - f(x)$, on obtient une estimation qui dépend des valeurs des dérivées $\frac{\partial f_i(\xi_i)}{\partial x_j}$ aux différents points ξ_i ($i = 1, 2, \ldots, n$) de l'intervalle $(x, x+\Delta x)$. L'inégalité (2′) montre qu'on peut se borner aux valeurs des dérivées $\frac{\partial f_i(\xi)}{\partial x_j}$ au même point $\xi \in (x, x+\Delta x)$.

T h é o r è m e. *Soit un vecteur fonction* $\varphi(x)$ *continu avec sa dérivée* $\varphi'(x)$ *dans un domaine fermé borné convexe* G *et*

$$\|\varphi'(x)\|_{II} \leqslant q < 1, \tag{3}$$

où q *est une constante. Si* $x^{(0)} \in G$ *et toutes les approximations successives*

$$x^{(p+1)} = \varphi(x^{(p)}) \quad (p = 0, 1, 2, \ldots) \tag{4}$$

sont aussi contenues dans G, *le processus itératif* (4) *converge vers une solution de l'équation*

$$x = \varphi(x) \tag{5}$$

qui est unique dans le domaine G.

D é m o n s t r a t i o n. Montrons que l'application $y = \varphi(x)$ est contractante dans G au sens de la l-norme.

Soit $x_1, x_2 \in G$ et $y_i = \varphi(x_i)$ $(i = 1, 2)$. En utilisant le lemme, on a:

$$\|y_1 - y_2\|_l = \|\varphi(x_1) - \varphi(x_2)\|_l \leqslant \|x_1 - x_2\|_l \cdot \|\varphi'(\xi)\|_l, \tag{6}$$

où $\xi \in G$.

Comme

$$\|\varphi'(\xi)\|_l \leqslant \max_{x \in G} \|\varphi'(x)\|_l = \|\varphi'(x)\|_{II} \leqslant q,$$

l'inégalité (6) entraîne:

$$\|y_1 - y_2\|_l \leqslant q \|x_1 - x_2\|_l,$$

avec $0 \leqslant q < 1$.

En vertu du théorème du § 10 le théorème est démontré.

C o r o l l a i r e. Le processus itératif (4) converge vers une solution de l'équation (5) et cette solution est unique si pour $x \in G$ les inégalités

$$\sum_{j=1}^{n} \left| \frac{\partial \varphi_j(x)}{\partial x_i} \right| \leqslant q_i < 1 \tag{7}$$

$(i = 1, 2, \ldots, n)$ sont respectées.

R e m a r q u e. Le théorème du § 10 entraîne pour l'approximation $x^{(p)}$ l'estimation suivante:

$$\|x^* - x^{(p)}\|_l \leqslant \frac{q^p}{1+q} \|x^{(1)} - x^{(0)}\|_l,$$

où $x^{(1)} = \varphi(x^{(0)})$.

§ 12*. Méthode de la plus grande pente (méthode du gradient)

Soit un système d'équations

$$\left.\begin{array}{l} f_1(x_1, x_2, \ldots, x_n)=0, \\ f_2(x_1, x_2, \ldots, x_n)=0, \\ \cdots\cdots\cdots\cdots \\ f_n(x_1, x_2, \ldots, x_n)=0 \end{array}\right\} \qquad (1)$$

ou sous forme vectorielle

$$f(x)=0 \qquad (2)$$

avec

$$f=\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}.$$

Supposons que les fonctions f_i sont réelles et continûment différentiables dans leur domaine de définition commun. Considérons la fonction

$$U(x)=\sum_{i=1}^{n}[f_i(x)]^2=(f(x), f(x)). \qquad (3)$$

Il est évident que chaque solution du système (1) annule la fonction $U(x)$; par contre, les nombres $x_1, x_2, \ldots, x_n$ pour lesquels la fonction $U(x)$ est nulle forment une solution du système (1).

Supposons que le système (1) n'ait qu'une solution isolée qui est le point du strict minimum de la fonction $U(x)$. Le problème se ramène alors à la recherche du minimum de la fonction $U(x)$ dans l'espace $E_n=\{x_1, x_2, \ldots, x_n\}$ de dimension n.

Soient x la solution du système (1) et $x^{(0)}$ son approximation initiale. Menons par le point $x^{(0)}$ la surface de niveau de la fonction $U(x)$. Si le point $x^{(0)}$ est suffisamment proche de la solution x, sous les hypothèses adoptées la surface de niveau

$$U(x) = U(x^{(0)})$$

ressemblera à un ellipsoïde.

Suivons à partir du point $x^{(0)}$ la normale à la surface $U(x) = U(x^{(0)})$ tant que cette normale ne touche en un certain point $x^{(1)}$ quelque autre surface de niveau (fig. 60)

$$U(x)=U(x^{(1)}).$$

Ensuite, en partant du point $x^{(1)}$, suivons encore la normale à la surface de niveau $U(x) = U(x^{(1)})$ tant que cette normale ne

touche en un certain point $x^{(2)}$ une nouvelle surface de niveau $U(x) = U(x^{(2)})$, etc.

Comme $U(x^{(0)}) > U(x^{(1)}) > U(x^{(2)}) > \ldots$, en nous déplaçant dans cette direction nous nous approchons rapidement du point à valeur minimale de U (fond de la « cuvette ») qui correspond à la solution cherchée x du système (1). Désignons par

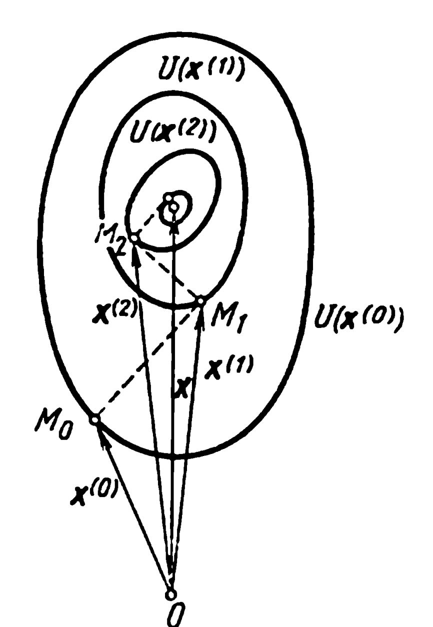


Fig. 60.

$$\operatorname{grad} U(x) = \begin{bmatrix} \frac{\partial U}{\partial x_1} \\ \vdots \\ \frac{\partial U}{\partial x_n} \end{bmatrix}$$

le gradient * de la fonction $U(x)$.

En considérant les triangles vectoriels OM_0M_1, OM_1M_2, ... on déduit que

$$x^{(p+1)} = x^{(p)} - \lambda_p \operatorname{grad} U(x^{(p)})$$
$$(p = 0, 1, 2, \ldots).$$

Il reste à déterminer les facteurs λ_p. A cette fin, considérons la fonction scalaire

$$\Phi(\lambda) = U[x^{(p)} - \lambda \operatorname{grad} U(x^{(p)})].$$

La fonction $\Phi(\lambda)$ donne la variation du niveau de la fonction U le long de la normale correspondante à la surface de niveau en $x^{(p)}$. Le facteur $\lambda = \lambda_p$ doit être choisi tel que $\Phi(\lambda)$ soit minimale. En dérivant par rapport à λ et en annulant la dérivée, on obtient l'équation

$$\Phi'(\lambda) = \frac{d}{d(\lambda)} U[x^{(p)} - \lambda \operatorname{grad} U(x^{(p)})] = 0. \tag{4}$$

La racine positive minimale de l'équation (4) nous donne précisément la valeur de λ_p. D'une façon générale, l'équation (4) doit être résolue numériquement. Nous indiquerons donc une méthode de calcul approchée des nombres λ_p. Considérons que λ est une petite valeur dont on peut négliger le carré et les puissances plus grandes.

---

* Le gradient de la fonction $U(x)$ (désigné par grad U ou ∇U; le symbole ∇ se lit *nabla*) est un vecteur appliqué au point x orienté suivant la normale n à la surface de niveau de la fonction au point donné dans le sens de croissance de U et de longueur égale à $\frac{\partial U}{\partial n}$.

La formule suivante a lieu

$$\operatorname{grad} U = \frac{\partial U}{\partial x_1} e_1 + \frac{\partial U}{\partial x_2} e_2 + \ldots + \frac{\partial U}{\partial x_n} e_n,$$

où e_i $(i = 1, 2, \ldots, n)$ sont des vecteurs unités de l'espace E_n.

On a

$$\Phi(\lambda) = \sum_{i=1}^{n} \{f_i [x^{(p)} - \lambda \operatorname{grad} U (x^{(p)})]\}^2.$$

La décomposition des fonctions f_i suivant les puissances de λ aux termes linéaires près donne

$$\Phi(\lambda) = \sum_{i=1}^{n} \left[f_i (x^{(p)}) - \lambda \frac{\partial f_i (x^{(p)})}{\partial x} \operatorname{grad} U (x^{(p)})\right]^2,$$

où

$$\frac{\partial f_i}{\partial x} = \left[\frac{\partial f_i}{\partial x_1}, \frac{\partial f_i}{\partial x_2}, \ldots, \frac{\partial f_i}{\partial x_n}\right].$$

D'où

$$\Phi'(\lambda) = -2 \sum_{i=1}^{n} \left[f_i (x^{(p)}) - \lambda \frac{\partial f_i (x^{(p)})}{\partial x} \operatorname{grad} U (x^{(p)})\right] \times$$

$$\times \frac{\partial f_i (x^{(p)})}{\partial x} \operatorname{grad} U (x^{(p)}) = 0.$$

Par conséquent

$$\lambda_p = \frac{\sum_{i=1}^{n} f_i (x^{(p)}) \frac{\partial f_i (x^{(p)})}{\partial x} \operatorname{grad} U (x^{(p)})}{\sum_{i=1}^{n} \left[\frac{\partial f_i (x^{(p)})}{\partial x} \operatorname{grad} U (x^{(p)})\right]^2} =$$

$$= \frac{(f (x^{(p)}), W (x^{(p)}) \operatorname{grad} U (x^{(p)}))}{(W (x^{(p)}) \operatorname{grad} U (x^{(p)}), W (x^{(p)}) \operatorname{grad} U (x^{(p)}))}$$

avec

$$W(x) = \frac{df}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \ldots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \ldots & \frac{\partial f_2}{\partial x_n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \ldots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

la matrice jacobienne du vecteur fonction f.

Ensuite, on a

$$\frac{\partial U}{\partial x_j} = \frac{\partial}{\partial x_j} \left\{\sum_{i=1}^{n} [f_i (x)]^2\right\} = 2 \sum_{i=1}^{n} f_i (x) \frac{\partial f_i (x)}{\partial x_j}.$$

D'où

$$\operatorname{grad} U(x) = 2\begin{bmatrix} \sum\limits_{i=1}^{n} \frac{\partial f_i(x)}{\partial x_1} f_i(x) \\ \cdots\cdots \\ \sum\limits_{i=1}^{n} \frac{\partial f_i(x)}{\partial x_n} f_i(x) \end{bmatrix} = 2W'(x) f(x),$$

avec $W'(x)$ la matrice jacobienne transposée.

Et finalement

$$\mu_p = 2\lambda_p = \frac{(f^{(p)},\ W_p W'_p f^{(p)})}{(W_p W'_p f^{(p)},\ W_p W'_p f^{(p)})}, \tag{5}$$

où, pour abréger l'écriture, nous avons posé

$$f^{(p)} = f(x^{(p)}); \quad W_p = W(x^{(p)}),$$

de plus

$$x^{(p+1)} = x^{(p)} - \mu_p W'_p f^{(p)} \quad (p = 0,\ 1,\ 2,\ \ldots). \tag{6}$$

Si l'on admet que la fonction $f(x)$ est deux fois continûment dérivable dans le voisinage de la solution cherchée x, on peut obtenir des formules de correction plus exactes $\Delta x^{(p)} = x^{(p+1)} - x^{(p)}$ (cf. [7]).

E x e m p l e. Calculer par la méthode de la plus grande pente les solutions approchées du système

$$\left.\begin{aligned} x + x^2 - 2yz &= 0{,}1; \\ y - y^2 + 3xz &= -0{,}2; \\ z + z^2 + 2xy &= 0{,}3 \end{aligned}\right\}$$

reposant dans le voisinage de l'origine des coordonnées.

S o l u t i o n. On a

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Ici

$$f = \begin{bmatrix} x + x^2 - 2yz - 0{,}1 \\ y - y^2 + 3xz + 0{,}2 \\ z + z^2 + 2xy - 0{,}3 \end{bmatrix}$$

et

$$W = \begin{bmatrix} 1 + 2x & -2z & -2y \\ 3z & 1 - 2y & 3x \\ 2y & 2x & 1 + 2z \end{bmatrix}.$$

En y portant l'approximation initiale, on aura :

$$f^{(0)}=\begin{bmatrix}-0,1\\ 0,2\\ -0,3\end{bmatrix} \quad \text{et} \quad W_0=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}=E.$$

Les formules (5) et (6) permettent d'obtenir la première approximation

$$\mu_0=\frac{(f^{(0)},\ f^{(0)})}{(f^{(0)},\ f^{(0)})}=1$$

et

$$x^{(1)}=x^{(0)}-1\cdot Ef^{(0)}=\begin{bmatrix}0,1\\ -0,2\\ 0,3\end{bmatrix}.$$

D'une façon analogue on obtient la deuxième approximation $x^{(1)}$. On a :

$$f^{(1)}=\begin{bmatrix}0,13\\ 0,05\\ 0,05\end{bmatrix}; \quad W_1=\begin{bmatrix}1,2&-0,6&0,4\\ 0,9&1,4&0,3\\ -0,4&0,2&1,6\end{bmatrix}.$$

D'où

$$W_1'f^{(1)}=\begin{bmatrix}0,181\\ 0,002\\ 0,147\end{bmatrix}$$

et

$$W_1W_1'f^{(1)}=\begin{bmatrix}0,2748\\ 0,2098\\ 0,1632\end{bmatrix}.$$

Par suite

$$\mu_1=\frac{0,13\cdot 0,2748+0,05\cdot 0,2098+0,05\cdot 0,1632}{0,2748^2+0,2098^2+0,1632^2}=\frac{0,054374}{0,14619797}=0,3719$$

et

$$x^{(2)}=\begin{bmatrix}0,1\\ -0,2\\ 0,3\end{bmatrix}-0,37119\cdot\begin{bmatrix}0,181\\ 0,002\\ 0,147\end{bmatrix}=\begin{bmatrix}0,0327\\ -0,2007\\ 0,2453\end{bmatrix}.$$

Pour vérifier, calculons le résidu

$$f^{(2)}=\begin{bmatrix}0,032\\ -0,017\\ -0,007\end{bmatrix}.$$

§ 13. Méthode de la plus grande pente pour le cas d'un système d'équations linéaires

Considérons un système d'équations linéaires

$$\left.\begin{aligned} f_1 &\equiv \sum_{j=1}^{n} a_{1j}x_j - b_1 = 0, \\ f_2 &\equiv \sum_{j=1}^{n} a_{2j}x_j - b_2 = 0, \\ &\cdots\cdots\cdots\cdots \\ f_n &\equiv \sum_{j=1}^{n} a_{nj}x_j - b_n = 0 \end{aligned}\right\} \tag{1}$$

à matrice réelle $A = [a_{ij}]$ et à colonne des termes constants

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Il vient

$$f = Ax - b$$

et

$$W = \frac{df}{dx} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} = A.$$

Par conséquent,

$$x^{(p+1)} = x^{(p)} - \mu_p A' r_p, \tag{2}$$

où $r_p = Ax^{(p)} - b$ est le *résidu* du vecteur $x^{(p)}$ et

$$\mu_p = \frac{(r_p,\ AA'r_p)}{(AA'r_p,\ AA'r_p)} \quad (p = 0,\ 1,\ 2,\ \dots) \tag{3}$$

(cf. [5], [6]).

L application des formules (2) et (3) conduit à des calculs très longs. Aussi en pratique, au lieu de « la plus grande pente » recourt-on à une « pente » simple en cherchant à minimiser la fonction

$$U = (Ax - b,\ Ax - b).$$

Dans ces conditions, le nombre de pas assurant la précision imposée des solutions du système (1) devient en général plus grand; par contre, on peut rendre le calcul de chaque pas plus simple.

Dans une formulation générale on adopte:

$$x^{(p+1)} = x^{(p)} - \lambda_p y^{(p)} \quad (p = 0, 1, 2, \ldots),$$

où $y^{(p)}$ est un vecteur arbitraire orienté vers l'extérieur de la surface de niveau U = const qui passe par le point $x^{(p)}$

$$(\mathrm{grad}\, U(x^{(p)}),\ y^{(p)}) > 0.$$

On a

$$r_{p+1} = Ax^{(p+1)} - b = Ax^{(p)} - b - \lambda_p A y^{(p)} = r_p - \lambda_p A y^{(p)}.$$

L'un des modes possibles déterminant le facteur scalaire est basé sur la restriction [7]

$$(r_{p+1},\ y^{(p)}) = (r_p,\ y^{(p)}) - \lambda_p (Ay^{(p)},\ y^{(p)}) = 0.$$

D'où

$$\lambda_p = \frac{(r_p,\ y^{(p)})}{(Ay^{(p)},\ y^{(p)})}.$$

Tel ou tel schéma de calcul s'obtient en fonction du choix du vecteur $y^{(p)}$. En particulier, si la matrice $A = A'$ est définie positive (chapitre X, § 15), en posant $y^{(p)} = r_p$ on aura:

$$x^{(p+1)} = x^{(p)} - \frac{(r_p,\ r_p)}{(Ar_p,\ r_p)} r_p$$

($p = 0, 1, 2, \ldots$), de plus $(\mathrm{grad}\, U(x^{(p)}),\ y^{(p)}) = 2(Ar_p,\ r_p) > 0$ pour $r_p \neq 0$.

E x e m p l e. Résoudre par la méthode de la plus grande pente le système d'équations

$$\left.\begin{aligned} 8x_1 - x_2 - 2x_3 &= 2{,}3\,; \\ 10x_2 + x_3 + 2x_4 &= -0{,}5\,; \\ -x_1 + 6x_3 + 2x_4 &= -1{,}2\,; \\ 3x_1 - x_2 + 2x_3 + 12x_4 &= 3{,}7. \end{aligned}\right\} \qquad (4)$$

S o l u t i o n. Comme dans la matrice du système prédominent les éléments diagonaux, on prend comme vecteur initial $x^{(0)}$ le vecteur dont les coordonnées sont des valeurs arrondies des solutions du système:

$$8x_1 = 2{,}3\,; \qquad 6x_3 = -1{,}2\,;$$
$$10x_2 = -0{,}5\,; \quad 12x_4 = 3{,}7.$$

Il en résulte, par exemple, que

$$x^{(0)} = \begin{bmatrix} 0{,}3 \\ -0{,}05 \\ -0{,}2 \\ 0{,}3 \end{bmatrix}.$$

Donc

$$\boldsymbol{r}_0 = Ax^{(0)} - \boldsymbol{b} =$$

$$= \begin{bmatrix} 8 & -1 & -2 & 0 \\ 0 & 10 & 1 & 2 \\ -1 & 0 & 6 & 2 \\ 3 & -1 & 2 & 12 \end{bmatrix} \begin{bmatrix} 0,3 \\ -0,05 \\ -0,2 \\ 0,3 \end{bmatrix} - \begin{bmatrix} 2,3 \\ -0,5 \\ -1,2 \\ 3,7 \end{bmatrix} = \begin{bmatrix} 0,55 \\ 0,4 \\ 0,3 \\ 0,45 \end{bmatrix}.$$

Ensuite

$$A'\boldsymbol{r}_0 = \begin{bmatrix} 8 & 0 & -1 & 3 \\ -1 & 10 & 0 & -1 \\ -2 & 1 & 6 & 2 \\ 0 & 2 & 2 & 12 \end{bmatrix} \begin{bmatrix} 0,55 \\ 0,4 \\ 0,3 \\ 0,45 \end{bmatrix} = \begin{bmatrix} 5,45 \\ 3,0 \\ 2,0 \\ 6,8 \end{bmatrix}$$

et

$$AA'\boldsymbol{r}_0 = \begin{bmatrix} 8 & -1 & -2 & 0 \\ 0 & 10 & 1 & 2 \\ -1 & 0 & 6 & 2 \\ 3 & -1 & 2 & 12 \end{bmatrix} \begin{bmatrix} 5,45 \\ 3,0 \\ 2,0 \\ 6,8 \end{bmatrix} = \begin{bmatrix} 36,6 \\ 45,6 \\ 20,15 \\ 98,95 \end{bmatrix}.$$

Appliquant la formule (3) on obtient :

$$\mu_0 = \frac{(\boldsymbol{r}_0,\ AA'\boldsymbol{r}_0)}{(AA'\boldsymbol{r}_0,\ AA'\boldsymbol{r}_0)} = \frac{0,55 \cdot 36,6 + 0,4 \cdot 45,6 + 0,3 \cdot 20,15 + 0,45 \cdot 98,95}{36,6^2 + 45,6^2 + 20,15^2 + 98,95^2} =$$

$$= \frac{88,9425}{13616,0452} = 0,006532.$$

D'où

$$x^{(1)} = x^{(0)} - \mu_0 A'\boldsymbol{r}_0 = \begin{bmatrix} 0,3 \\ -0,05 \\ -0,2 \\ 0,3 \end{bmatrix} - 0,006532 \begin{bmatrix} 5,45 \\ 3,0 \\ 2,0 \\ 6,8 \end{bmatrix} = \begin{bmatrix} 0,2644 \\ -0,0696 \\ -0,2131 \\ 0,2556 \end{bmatrix},$$

de plus,

$$\boldsymbol{r}^{(1)} = Ax^{(1)} - \boldsymbol{b} = \begin{bmatrix} 0,3109 \\ 0,1020 \\ 0,1684 \\ -0,1966 \end{bmatrix}.$$

D'une façon analogue on trouve les approximations ultérieures et les résidus correspondants :

$$x^{(2)} = \begin{bmatrix} 0{,}2351 \\ -0{,}0849 \\ -0{,}2147 \\ 0{,}2863 \end{bmatrix}, \quad r_2 = \begin{bmatrix} 0{,}0956 \\ 0{,}0087 \\ 0{,}2493 \\ 0{,}0967 \end{bmatrix};$$

$$x^{(3)} = \begin{bmatrix} 0{,}2296 \\ -0{,}0842 \\ -0{,}2251 \\ 0{,}2748 \end{bmatrix}, \quad r_3 = \begin{bmatrix} 0{,}0712 \\ -0{,}0280 \\ 0{,}1692 \\ -0{,}0806 \end{bmatrix};$$

$$x^{(4)} = \begin{bmatrix} 0{,}2266 \\ -0{,}0792 \\ -0{,}2379 \\ 0{,}2875 \end{bmatrix}, \quad r_4 = \begin{bmatrix} 0{,}0680 \\ 0{,}0354 \\ 0{,}1211 \\ 0{,}0334 \end{bmatrix};$$

$$x^{(5)} = \begin{bmatrix} 0{,}2228 \\ -0{,}0810 \\ -0{,}2430 \\ 0{,}2823 \end{bmatrix}, \quad r_5 = \begin{bmatrix} 0{,}0493 \\ 0{,}0013 \\ 0{,}0839 \\ -0{,}0493 \end{bmatrix},$$

etc.

Remarquons que dans le cas considéré le processus des approximations converge lentement ; après la cinquième approximation nous sommes encore loin de la solution exacte du système (4), celle-ci étant $x_1 = 0{,}2$; $x_2 = -0{,}1$; $x_3 = -0{,}3$; $x_4 = 0{,}3$.

§ 14*. Méthode des séries entières

Soit un système non linéaire

$$f_k(x_1, x_2, \ldots, x_n) = 0 \tag{1}$$

($k = 1, 2, \ldots, n$), où les fonctions f_k sont analytiques dans le voisinage d'une solution isolée $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$.

Considérons un système plus général [8]

$$F_k(x_1, x_2, \ldots, x_n; \quad \lambda) = 0 \tag{2}$$

($k = 1, 2, \ldots, n$), qui dépend du paramètre réel λ et tel que pour $\lambda = 0$ la solution de (2) est immédiate, alors que pour $\lambda = 1$ le système (2) est identique au système (1)

$$F_k(x_1, x_2, \ldots, x_n; 1) \equiv f_k(x_1, x_2, \ldots, x_n)$$

($k = 1, 2, \ldots, n$). Le paramètre λ doit être introduit de façon que la relation des fonctions F_k par rapport à λ soit la plus simple possible. Par exemple, si $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ est une approximation grossière de la solution, on peut poser

$$\sum_{j=1}^{n} (x_j - x_j^{(0)}) \frac{\partial f_k(x^{(0)})}{\partial x_j} + \lambda \left[f_k(x) - \sum_{j=1}^{n} (x_j - x_j^{(0)}) \frac{\partial f_k(x^{(0)})}{\partial x_j} \right] = 0$$

($k = 1, 2, \ldots, n$), où

$$x = (x_1, x_2, \ldots, x_n).$$

Nous supposerons que les F_k soient des fonctions analytiques de λ pour $|\lambda| \leqslant 1$.

Supposons que pour $|\lambda| \leqslant 1$ le système (2) admette une solution analytique simple $x_j(\lambda)$ ($j = 1, 2, \ldots, n$) coïncidant pour $\lambda = 1$ avec x_j^* ($j = 1, 2, \ldots, n$). Posons

$$x_j(0) = x_j^{(0)} \quad (j = 1, 2, \ldots, n),$$

où $x_j^{(0)}$ ($j = 1, 2, \ldots, n$) est une solution connue du système (2) pour $\lambda = 0$. En développant les fonctions $x_j(\lambda)$ en série de Taylor au point $\lambda = 0$, on obtient:

$$x_j(\lambda) = x_j(0) + \lambda x_j'(0) + \frac{\lambda^2}{2!} x_j''(0) + \ldots \quad (j = 1, 2, \ldots, n). \quad (3)$$

Pour calculer les coefficients $x_j'(0)$ dérivons l'égalité (2) par rapport au paramètre λ:

$$\sum_{j=1}^{n} \frac{\partial F_k}{\partial x_j} x_j'(\lambda) + \frac{\partial F_k}{\partial \lambda} = 0 \quad (k = 1, 2, \ldots, n). \quad (4)$$

Posant $x = x^{(0)}$ et $\lambda = 0$, on aura

$$\sum_{j=1}^{n} \frac{\partial F_k(x^{(0)}; 0)}{\partial x_j} x_j'(0) = -\frac{\partial F_k(x^{(0)}; 0)}{\partial \lambda} \quad (k = 1, 2, \ldots, n).$$

Si

$$\det \left[\frac{\partial F_k(x^{(0)}; 0)}{\partial x_j} \right] \neq 0,$$

on trouve $x_j'(0)$.

Ensuite en dérivant par rapport à λ l'égalité (4), on obtient:

$$\sum_{j=1}^{n} \frac{\partial F_k}{\partial x_j} x_j''(\lambda) + \sum_{j=1}^{n} \sum_{l=1}^{n} \frac{\partial^2 F_k}{\partial x_j \partial x_l} x_j'(\lambda) x_l'(\lambda) +$$
$$+ 2 \sum_{j=1}^{n} \frac{\partial^2 F_k}{\partial x_j \partial \lambda} x_j'(\lambda) + \frac{\partial^2 F_k}{\partial \lambda^2} = 0.$$

D'où pour $x = x^{(0)}$ et $\lambda = 0$, on trouve :

$$\sum_{j=1}^{n} \frac{\partial F_k(x^{(0)};\ 0)}{\partial x_j} x_j''(0) = -\sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 F_k(x^{(0)};\ 0)}{\partial x_j\, \partial x_i} x_j'(0)\ x_i'(0) -$$

$$-2 \sum_{j=1}^{n} \frac{\partial^2 F_k(x^{(0)};\ 0)}{\partial x_j\, \partial \lambda} x_j'(0) - \frac{\partial^2 F_k(x^{(0)};\ 0)}{\partial \lambda^2} \quad (k = 1,\ 2,\ \ldots,\ n). \quad (5)$$

Comme les $x_j'(0)$ sont connues, le système (5) permet de déterminer $x_j''(0)$. D'une façon analogue on calcule les dérivées $x'''(0)$, $x^{\mathrm{IV}}(0)$, . . .

Remarquons que la matrice des coefficients des dérivées supérieures est toujours la même et est égale à la matrice jacobienne des fonctions $F_1, F_2, \ldots, F_n$ relativement aux variables $x_1, x_2, \ldots, x_n$ pour $x_j = x_j^{(0)}$ $(j = 1, 2, \ldots, n)$ et $\lambda = 0$.

En supposant que les séries (3) convergent pour $\lambda = 1$, on obtient finalement :

$$x_j^* = x_j(1) = x_j(0) + x_j'(0) + \frac{1}{2!} x_j''(0) + \ldots \quad (j = 1,\ 2,\ \ldots,\ n). \quad (6)$$

L'inconvénient de la méthode est dû au calcul compliqué dans le cas général des dérivées d'ordres supérieurs. D'autre part, la rapidité de la convergence de la série (6) peut être insuffisante.

L'application de la méthode n'impose pas nécessairement aux fonctions $x_j(\lambda)$ $(j = 1, 2, \ldots, n)$ d'être analytiques, à savoir: au lieu de la série de Taylor on peut faire appel à la formule de Taylor en interrompant les séries $x_j(\lambda)$ à une certaine puissance λ^s et en évaluant leurs restes d'après les formules connues (chapitre III, § 4).

BIBLIOGRAPHIE

1. *L. Kantorovitch.* Sur la méthode de Newton. Travaux de l'institut des mathématiques Stéklov, XXVIII. Moscou-Léningrad, 1949, pp. 104-144.
2. *A. Ostrowski.* Recueil de travaux à la mémoire de D. A. Grave. 1940, p. 213.
3. *J. B. Scarborough.* Numerical Mathematical Analysis. John Hopkins, 1950, chapitre IX.
4. *D. Ventsel, E. Ventsel.* Eléments de la théorie des calculs approchés. Editions de l'Académie militaire technique de l'Air Joukovski, Moscou, 1949, chapitre III, § 8.
5. *W. E. Milne.* Numerical solutions of differential equations.
6. *A. S. Housholder.* Principles of Numerical Analysis. Mc Graw-Hill, 1953, chapitre III.
7. *A. D. Booth.* Numerical methods. London, Butterworth, 1955.
8. Modern Mathematics for the Engineer, sous la direction d'*E. F. Beckenbach.* Mc Graw-Hill, 1956, chapitre XIV. C. B. Morrey Jr. Non linear methods.

CHAPITRE XIV

INTERPOLATION DES FONCTIONS

§ 1. Différences finies successives

Soit

$$y = f(x)$$

la fonction donnée. Désignons par $\Delta x = h$ une valeur fixée de l'accroissement de l'argument (*pas*). Alors l'expression

$$\Delta y \equiv \Delta f(x) = f(x + \Delta x) - f(x) \qquad (1)$$

s'appelle *différence première* de la fonction y. D'une façon analogue on définit les *différences d'ordres supérieurs*

$$\Delta^n y = \Delta(\Delta^{n-1} y) \quad (n = 2, 3, \ldots).$$

Par exemple,

$$\Delta^2 y = \Delta[f(x + \Delta x) - f(x)] =$$
$$= [f(x + 2\Delta x) - f(x + \Delta x)] - [f(x + \Delta x) - f(x)] =$$
$$= f(x + 2\Delta x) - 2f(x + \Delta x) + f(x).$$

Exemple. Construire les différences de la fonction

$$P(x) = x^3$$

en considérant le pas $\Delta x = 1$.

Solution. On a:

$$\Delta P(x) = (x + 1)^3 - x^3 = 3x^2 + 3x + 1,$$
$$\Delta^2 P(x) = [3(x + 1)^2 + 3(x + 1) + 1] -$$
$$- (3x^2 + 3x + 1) = 6x + 6,$$
$$\Delta^3 P(x) = [6(x + 1) + 6] - (6x + 6) = 6,$$
$$\Delta^n P(x) = 0 \quad \text{pour} \quad n > 3.$$

Il est à remarquer que la différence troisième de la fonction $P(x)$ est constante.

Dans le cas général la proposition suivante est vraie: si

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

est un polynôme de degré n, alors $\Delta^n P_n(x) = n!a_0h^n = \text{const}$, où $\Delta x = h$.

En effet, on a:

$$\Delta P_n(x) = P_n(x+h) - P_n(x) =$$
$$= a_0[(x+h)^n - x^n] + a_1[(x+h)^{n-1} - x^{n-1}] + \ldots$$
$$\ldots + a_{n-1}[(x+h) - h].$$

En se débarrassant des parenthèses suivant le binôme de Newton, on voit aisément que $\Delta P_n(x)$ est un polynôme de degré $(n-1)$:

$$\Delta P_n(x) = b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}$$

où

$$b_0 = nha_0.$$

Des raisonnements analogues conduisent à la conclusion que la différence seconde $\Delta^2 P_n(x)$ est un polynôme de degré $(n-2)$:

$$\Delta^2 P_n(x) = c_0x^{n-2} + c_1x^{n-3} + \ldots + c_{n-2}$$

et

$$c_0 = (n-1)hb_0 = n(n-1)h^2a_0.$$

En raisonnant ainsi on établit finalement, de proche en proche, que

$$\Delta^n P_n(x) = n!a_0h^n = \text{const},$$

d'où l'on tire en conséquence

$$\Delta^s P_n(x) = 0 \quad \text{pour} \quad s > n.$$

Le symbole Δ (delta) peut être considéré comme un *opérateur* qui associe à la fonction $y = f(x)$ la fonction $\Delta y = f(x + \Delta x) - f(x)$ (Δx étant une constante). Les propriétés principales d'un opérateur Δ se vérifient facilement:

1) $\Delta(u + v) = \Delta u + \Delta v$;

2) $\Delta(Cu) = C\Delta u$ (C est une constante);

3) $\Delta^m(\Delta^n y) = \Delta^{m+n} y$,

où m et n sont des entiers non négatifs; par définition on pose $\Delta^0 y = y$.

La formule (1) entraîne

$$f(x + \Delta x) = f(x) + \Delta f(x);$$

en considérant Δ comme un facteur symbolique, on obtient:

$$f(x + \Delta x) = (1 + \Delta) f(x). \tag{2}$$

En appliquant cette relation n fois de suite on aura:

$$f(x + n\Delta x) = (1 + \Delta)^n f(x). \tag{3}$$

L'utilisation de la formule du binôme de Newton * conduit finalement à

$$f(x+n\Delta x)=\sum_{m=0}^{n} C_n^m \Delta^m f(x), \tag{4}$$

avec

$$C_n^m=\frac{n(n-1)\ldots[n-(m-1)]}{m!}$$

le nombre de combinaisons de n éléments pris m à m.

Ainsi la formule (4) permet d'exprimer les valeurs successives de la fonction $f(x)$ par ses différences de divers ordres.

En recourant à l'identité

$$\Delta=(1+\Delta)-1 \tag{5}$$

et en appliquant le binôme de Newton, on obtient:

$$\Delta^n f(x)=[(1+\Delta)-1]^n f(x)=(1+\Delta)^n f(x)-C_n^1(1+\Delta)^{n-1} f(x)+$$
$$+C_n^2(1+\Delta)^{n-2} f(x)-\ldots+(-1)^n f(x).$$

Il en résulte en vertu de la formule (3):

$$\Delta^n f(x)=f(x+n\Delta x)-C_n^1 f[x+(n-1)\Delta x]+$$
$$+C_n^2 f[x+(n-2)\Delta x]-\ldots+(-1)^n f(x). \tag{6}$$

La formule (6) exprime la différence d'ordre n de la fonction $f(x)$ par les valeurs successives de cette fonction.

Supposons que la fonction $f(x)$ ait une dérivée continue $f^{(n)}(x)$ sur le segment $[x, x+n\Delta x]$. On a alors la formule importante

$$\Delta^n f(x)=(\Delta x)^n f^{(n)}(x+\theta n\Delta x), \tag{7}$$

avec

$$0<\theta<1.$$

Pour démontrer la formule (7), le plus simple est de le faire par récurrence.

En effet, avec $n=1$ on obtient le théorème de Lagrange des accroissements finis et, par conséquent, la formule (7) est vraie. Supposons maintenant qu'on ait pour $k<n$:

$$\Delta^k f(x)=(\Delta x)^k f^{(k)}(x+\theta' k\Delta x),$$

avec

$$0<\theta'<1.$$

Il vient

$$\Delta^{k+1} f(x)=\Delta^k [f(x+\Delta x)-f(x)]=$$
$$=(\Delta x)^k [f^{(k)}(x+\Delta x+\theta' k\Delta x)-f^{(k)}(x+\theta' k\Delta x)].$$

* Nous laissons au lecteur le soin de justifier l'application de la formule du binôme de Newton.

Appliquant le théorème de Lagrange à l'accroissement obtenu de la dérivée $f^{(k)}(x)$, on aura:

$$\Delta^{k+1} f(x) = (\Delta x)^k \Delta x f^{(k+1)}(x + \theta' k\Delta x + \theta'' \Delta x),$$

où $0 < \theta'' < 1$. Posant

$$\frac{\theta' k + \theta''}{k+1} = \theta, \tag{8}$$

on obtient finalement

$$\Delta^{k+1} f(x) = (\Delta x)^{k+1} f^{(k+1)}(x + \theta(k+1)\Delta x),$$

de plus, il est évident que

$$0 < \theta < 1.$$

Nous avons établi de cette façon le passage de k à $k+1$ et la formule (7) est ainsi démontrée.

La formule (7) entraîne

$$f^{(n)}(x + \theta n\Delta x) = \frac{\Delta^n f(x)}{(\Delta x)^n}.$$

D'où, en passant à la limite lorsque $\Delta x \to 0$ et en supposant que la dérivée $f^{(n)}(x)$ soit continue,

$$f^{(n)}(x) = \lim_{\Delta x \to 0} \frac{\Delta^n f(x)}{(\Delta x)^n}. \tag{9}$$

Il s'ensuit que pour des Δx petits la formule approchée

$$f^{(n)}(x) \approx \frac{\Delta^n f(x)}{(\Delta x)^n} \tag{10}$$

est vraie.

§ 2. Table des différences

Il arrive souvent que les fonctions $y = f(x)$ à étudier sont données par des valeurs tabulées $y_i = f(x_i)$ pour un système de points équidistants x_i $(i = 0, 1, 2, \ldots)$, où

$$\Delta x_i = x_{i+1} - x_i = h = \text{const}.$$

Les différences de la suite y_i sont déterminées naturellement par les relations

$$\Delta y_i = y_{i+1} - y_i,$$
$$\Delta^2 y_i = \Delta(\Delta y_i) = \Delta y_{i+1} - \Delta y_i,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\Delta^n y_i = \Delta(\Delta^{n-1} y_i) = \Delta^{n-1} y_{i+1} - \Delta^{n-1} y_i.$$

La première égalité entraîne

$$y_{i+1} = y_i + \Delta y_i = (1 + \Delta) y_i.$$

On en déduit successivement:

$$y_{i+2} = (1 + \Delta)\, y_{i+1} = (1 + \Delta)^2\, y_i,$$
$$y_{i+3} = (1 + \Delta)\, y_{i+2} = (1 + \Delta)^3\, y_i,$$
$$\dots\dots\dots\dots\dots\dots$$
$$y_{i+n} = (1 + \Delta)^n\, y_i.$$

En utilisant le binôme de Newton, on obtient:

$$y_{i+n} = y_i + C_n^1 \Delta y_i + C_n^2 \Delta^2 y_i + \dots + \Delta^n y_i.$$

Inversement, on a:

$$\Delta^n y_i = [(1+\Delta) - 1]^n\, y_i = (1+\Delta)^n\, y_i - C_n^1 (1+\Delta)^{n-1}\, y_i + \\ + C_n^2 (1+\Delta)^{n-2} y_i - \dots + (-1)^n\, y_i$$

ou

$$\Delta^n y_i = y_{n+i} - C_n^1 y_{n+i-1} + C_n^2 y_{n+i-2} - \dots + (-1)^n\, y_i.$$

Par exemple

$$\Delta^2 y_i = y_{i+2} - 2y_{i+1} + y_i,$$
$$\Delta^3 y_i = y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i,$$

etc. Constatons que pour calculer la différence d'ordre n de $\Delta^n y_i$, il faut connaître $n + 1$ termes $y_i, y_{i+1}, \dots, y_{i+n}$ de la suite donnée.

Il est commode de ranger les différences finies successives dans les tableaux de deux types: *horizontal* (tableau 33) et *diagonal* (tableau 34).

Tableau 33

Tableau des différences horizontal

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
x_0	y_0	Δy_0	$\Delta^2 y_0$	$\Delta^3 y_0$
x_1	y_1	Δy_1	$\Delta^2 y_1$	$\Delta^3 y_1$
x_2	y_2	Δy_2	$\Delta^2 y_2$	$\Delta^3 y_2$
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .

Tableau 34

Tableau des différences diagonal

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
x_0	y_0			
		Δy_0		
x_1	y_1		$\Delta^2 y_0$	$\Delta^3 y_0$
		Δy_1		
x_2	y_2		$\Delta^2 y_1$	
		Δy_2		
x_3	y_3			

E x e m p l e 1. Former le tableau horizontal de la fonction

$$y = 2x^3 - 2x^2 + 3x - 1 \qquad (1)$$

à partir de la valeur initiale $x_0 = 0$, en adoptant le pas $h = 1$.

S o l u t i o n. En posant $x_0 = 0$, $x_1 = 1$ $x_2 = 2$, on trouve les valeurs correspondantes $y_0 = -1$, $y_1 = 2$, $y_2 = 13$. On en tire:

$$\Delta y_0 = y_1 - y_0 = 3,$$
$$\Delta y_1 = y_2 - y_1 = 11,$$
$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = 8.$$

Portons ces valeurs sur le tableau 35. Notre fonction étant un polynôme de troisième degré, sa différence troisième est constante (cf. § 1) et égale à

$$\Delta^3 y_i = 2 \cdot 3! = 12.$$

Pour poursuivre la formation du tableau 35 on peut donc recourir à la sommation en utilisant les formules

$$\Delta^2 y_{i+1} = \Delta^2 y_i + 12 \qquad (i = 0, 1, 2, \ldots),$$
$$\Delta y_{i+1} = \Delta y_i + \Delta^2 y_i \qquad (i = 1, 2, \ldots),$$
$$y_{i+1} = y_i + \Delta y_i \qquad (i = 2, 3, \ldots).$$

La ligne étagée indique les données initiales pour former le tableau.

Tableau 35

Tableau des différences horizontal de la fonction du troisième degré

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0	−1	3	8	12
1	2	11	20	12
2	13	31	32	12
3	44	63	44	12
4	107	107	56	12
5	214	163	68	12
. . .	. . .	. . .	. . .	. . .

R e m a r q u e. En composant le tableau des différences le calculateur peut commettre des erreurs aléatoires. Voyons quelle sera l'influence qu'exerce l'erreur ε de y_n sur les valeurs des différences. Le tableau des différences diagonal correspondant s'obtient sous forme de tableau 36 qui montre que: 1) si y_n est entaché d'une

Tableau 36

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
⋮	⋮	⋮	⋮	⋮	⋮
x_{n-4}	y_{n-4}				
		Δy_{n-4}			
x_{n-3}	y_{n-3}		$\Delta^2 y_{n-4}$		
		Δy_{n-3}		$\Delta^3 y_{n-4}$	
x_{n-2}	y_{n-2}		$\Delta^2 y_{n-3}$		$\Delta^4 y_{n-4}+\varepsilon$
		Δy_{n-2}		$\Delta^3 y_{n-3}+\varepsilon$	
x_{n-1}	y_{n-1}		$\Delta^2 y_{n-2}+\varepsilon$		$\Delta^4 y_{n-3}-4\varepsilon$
		$\Delta y_{n-1}+\varepsilon$		$\Delta^3 y_{n-2}-3\varepsilon$	
x_n	$y_n+\varepsilon$		$\Delta^2 y_{n-1}-2\varepsilon$		$\Delta^4 y_{n-2}+6\varepsilon$
		$\Delta y_n-\varepsilon$		$\Delta^3 y_{n-1}+3\varepsilon$	
x_{n+1}	y_{n+1}		$\Delta^2 y_n+\varepsilon$		$\Delta^4 y_{n-1}-4\varepsilon$
		Δy_{n+1}		$\Delta^3 y_n-\varepsilon$	
x_{n+2}	y_{n+2}		$\Delta^2 y_{n+1}$		$\Delta^4 y_n+\varepsilon$
		Δy_{n+2}		$\Delta^3 y_{n+1}$	
x_{n+3}	y_{n+3}		$\Delta^2 y_{n+2}$		
		Δy_{n+3}			
x_{n+4}	y_{n+4}				

erreur, les différences

$$\Delta y_{n-1}, \quad \Delta y_n; \quad \Delta^2 y_{n-2}, \quad \Delta^2 y_{n-1}, \quad \Delta^2 y_n,$$

etc., sont également fausses; 2) les erreurs entrent dans les $\Delta^k y$ différences d'ordre k avec des coefficients binomiaux aux signes alternés; plus précisément les valeurs respectives des erreurs sont

$$C_k^0\,\varepsilon, \quad -C_k^1\,\varepsilon, \quad C_k^2\,\varepsilon, \quad \ldots, \quad (-1)^k\,C_k^k\varepsilon$$

et, par conséquent, la valeur absolue de l'erreur maximale de la différence d'ordre k croît rapidement avec le numéro de cette diffé-

rence ; 3) pour toute différence $\Delta^k y$ la somme des erreurs, compte tenu de leurs signes, est nulle, alors que la somme des valeurs absolues des erreurs est $|\varepsilon|\cdot 2^k$. Ainsi, même une erreur négligeable de la valeur de la fonction conduit à des erreurs importantes dans ses différences d'ordres élevés. Remarquons que dans le cas d'un tableau diagonal l'erreur maximale des différences $\Delta^k y$ se trouve sur la même ligne horizontale que la valeur tabulée erronée y_n ou sur les lignes supérieure et inférieure voisines.

La *loi* examinée *de la propagation de l'erreur* ε dans le tableau des différences permet parfois d'établir l'existence et l'emplacement de cette erreur, ainsi que sa valeur numérique, et par là corriger le tableau.

Les tableaux des différences se composent en général à une unité décimale fixée près. Si la fonction $y = f(x)$ possède des dérivées continues jusqu à l'ordre m, le pas $h = \Delta x$ étant suffisamment petit, ses différences changent régulièrement jusqu'à l'ordre m y compris, la différence d'ordre m étant presque constante dans les limites des décimales données. Si cette dernière condition est enfreinte dans quelque partie du tableau et si la fonction n'a rien de singulier, on est alors en présence d'une erreur de calcul.

Après avoir trouvé l'écart maximal entre la différence d'ordre m et l'allure régulière, on peut déterminer l'emplacement de cette erreur dans la colonne des valeurs de la fonction y sous l'hypothèse que: 1) cette erreur est unique et résulte du calcul erroné d'une valeur de la fonction et 2) le calcul des différences finies n'a pas donné lieu à d'autres erreurs. Si on découvre une telle erreur dans le tableau des différences, on peut la corriger à l'aide das valeurs des différences. Montrons comment on le fait tout en nous bornant, pour simplifier, au cas des différences constantes secondes ou troisièmes.

Supposons que la valeur tabulée fausse est $y_n + \varepsilon$, où l'indice n est établi, alors que la valeur de l'erreur ε est inconnue.

Si les différences troisièmes sont pratiquement constantes, les différences secondes forment une progression arithmétique ; la valeur exacte de la différence seconde $\Delta^2 y_{n-1}$ sera donc égale à la moyenne arithmétique de trois différences fausses adjacentes:

$$\Delta^2 y_{n-1} = \frac{1}{3}[(\Delta^2 y_{n-2} + \varepsilon) + (\Delta^2 y_{n-1} - 2\varepsilon) + (\Delta^2 y_n + \varepsilon)],$$

du fait que les termes en ε se compensent.

D'après la valeur exacte de la différence seconde $\Delta^2 y_{n-1}$ on peut calculer la valeur de l'erreur ε, à savoir: cette erreur sera égale à la demi-différence des valeurs corrigée et fausse de la différence $\Delta^2 y_{n-1}$

$$\varepsilon = \frac{1}{2}[\Delta^2 y_{n-1} - (\Delta^2 y_{n-1} - 2\varepsilon)].$$

Quant à la valeur exacte de la fonction y_n elle-même, on l'obtient de l'identité

$$y_n = (y_n + \varepsilon) - \varepsilon.$$

Pour vérifier il faut de nouveau calculer les différences.

Exemple 2. Corriger l'erreur du tableau 37.

Tableau 37

Différences à erreur unique

x	y	Δy	$\Delta^2 y$	Erreur
15	13,260			
		884		
16	14,144		0	
		884		
17	15,028		0	
		884		ε
18	15,912		(−4) 0	
		88 (0) 4		
19	16,79 (2) 6		(8) 0	-2ε
		88 (8) 4		
20	17,680		(−4) 0	ε
		884		
21	18,564		0	
		884		
22	19,448		0	
		884		
23	20,332			

Solution. La perturbation maximale du changement régulier des différences secondes a lieu pour $x = 19$. L'erreur concerne trois lignes réunies par une accolade. Cherchons à déterminer la moyenne arithmétique de la différence seconde pour la ligne médiane des trois lignes associées

$$\Delta^2 y_{n-1} = \frac{10^{-3}}{3}(-4+8-4) = 0.$$

D'où

$$\varepsilon = \frac{1}{2}[0-0{,}008] = -0{,}004.$$

En corrigeant la valeur tabulée de y pour $x = 19$, on obtient:

$$y_n = (y_n + \varepsilon) - \varepsilon = 16{,}792 - (-0{,}004) = 16{,}796.$$

Après la correction on a un tableau dans lequel les différences premières changent régulièrement et la différence seconde est constante (les chiffres incorrects sont mis entre parenthèses). Notons que cette méthode ne permet de corriger que des erreurs de calcul isolées ou des

lapsus. Pour éliminer un grand nombre d'erreurs dues à des causes différentes, ainsi que pour réduire la cumulation des erreurs produites par le manque de précision des méthodes numériques elles-mêmes et l'arrondissement des résultats intermédiaires jusqu'au nombre de chiffres donné, on emploie des procédés de « lissage » spéciaux [1].

§ 3. Puissance généralisée

Dans ce qui suit nous devrons recourir à la notion de *puissance généralisée* [1].

D é f i n i t i o n. On appelle puissance généralisée n-ième du nombre x le produit de n facteurs dont le premier est égal à x et chaque facteur suivant est plus petit de h que le précédent:

$$x^{[n]} = x(x-h)(x-2h)\dots[x-(n-1)h], \tag{1}$$

où h est une constante fixée.

L'exposant d'une puissance généralisée se met généralement entre crochets. On pose $x^{[0]} = 1$.

Pour $h = 0$, la puissance généralisée (1) coïncide avec la puissance ordinaire

$$x^{[n]} = x^n.$$

Calculons les différences d'une puissance généralisée en posant $\Delta x = h$. Pour la différence première on a:

$$\begin{aligned}\Delta x^{[n]} &= (x+h)^{[n]} - x^{[n]} = \\ &= (x+h)x\dots[x-(n-2)h] - x(x-h)\dots[x-(n-1)h] = \\ &= x(x-h)\dots[x-(n-2)h]\cdot\{(x+h)-[x-(n-1)h]\} = \\ &= x(x-h)\dots[x-(n-2)h]\,nh = nhx^{[n-1]},\end{aligned}$$

soit

$$\Delta x^{[n]} = nhx^{[n-1]}. \tag{2}$$

Calculons la différence seconde:

$$\begin{aligned}\Delta^2 x^{[n]} &= \Delta(\Delta x^{[n]}) = \Delta(nhx^{[n-1]}) = \\ &= nh\cdot(n-1)hx^{[n-2]} = nh^2(n-1)x^{[n-2]}.\end{aligned}$$

Ainsi

$$\Delta^2 x^{[n]} = n(n-1)h^2 x^{[n-2]}.$$

Il est facile de déduire par récurrence la formule générale

$$\Delta^k x^{[n]} = n(n-1)\dots[n-(k-1)]\,h^k x^{[n-k]},$$

où $k = 1, 2, \dots, n$.

Il est clair que

$$\Delta^k x^{[n]} = 0 \quad \text{pour} \quad k > n.$$

La formule (2) permet également de déduire une formule simple de *sommation finie*. Soient

$$x_0,\ x_1,\ x_2,\ \ldots$$

des points équidistants de pas h

$$x_{i+1}-x_i=h \quad (i=0,\ 1,\ 2,\ \ldots).$$

Considérons la somme

$$S_N=\sum_{i=0}^{N-1} x_i^{[n]}.$$

Comme en vertu de la formule (2) on a:

$$x^{[n]}=\frac{\Delta x^{[n+1]}}{h(n+1)},$$

il vient

$$S_N=\frac{1}{h(n+1)}\sum_{i=0}^{N-1}\Delta x_i^{[n+1]}=$$

$$=\frac{1}{h(n+1)}\{x_1^{[n+1]}-x_0^{[n+1]}+x_2^{[n+1]}-x_1^{[n+1]}+\ldots+x_N^{[n+1]}-x_{N-1}^{[n+1]}\}=$$

$$=\frac{1}{h(n+1)}(x_N^{[n+1]}-x_0^{[n+1]}).$$

Ainsi

$$\sum_{i=0}^{N-1} x_i^{[n]}=\frac{x_N^{[n+1]}-x_0^{[n+1]}}{h(n+1)}. \qquad (3)$$

La formule (3) est analogue à la formule de Newton-Leibniz pour une puissance positive entière.

§ 4. Position du problème d'interpolation

Le problème d'interpolation le plus simple [2] consiste en ce qui suit. On donne sur le segment $[a, b]$ $n+1$ points $x_0, x_1, \ldots, x_n$ qui s'appellent *points d'interpolation*, et les valeurs d'une certaine fonction $f(x)$ en ces points

$$f(x_0)=y_0,\ f(x_1)=y_1,\ \ldots,\ f(x_n)=y_n. \qquad (1)$$

Soit à former la fonction $F(x)$ (*fonction d'interpolation*) qui appartient à une certaine classe connue et qui prend aux points d'interpolation les mêmes valeurs que $f(x)$, c'est-à-dire telle que

$$F(x_0)=y_0,\ \ F(x_1)=y_1,\ \ldots,\ F(x_n)=y_n. \qquad (2)$$

Géométriquement cela signifie qu'il faut trouver une courbe d'équation $y = F(x)$ et de type donné passant par le système des points donné $M_i(x_i, y_i)$ $(i = 0, 1, 2, \ldots)$ (fig. 61).

Le problème posé sous une forme générale peut avoir un nombre infini de solutions ou n'en avoir point. Toutefois, il a une et une

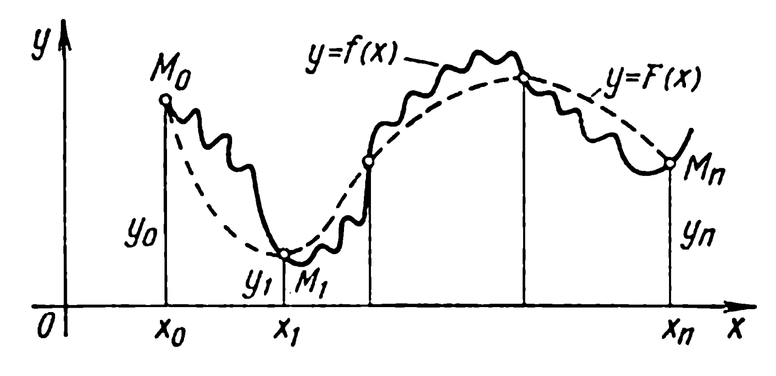


Fig. 61.

seule solution si l'on cherche non pas une fonction arbitraire $F(x)$ mais un polynôme $P_n(x)$ de degré inférieur ou égal à n vérifiant les conditions (2) et tel que

$$P_n(x_0) = y_0, \quad P_n(x_1) = y_1, \quad \ldots, \quad P_n(x_n) = y_n.$$

La formule d'interpolation obtenue

$$y = F(x)$$

s'emploie généralement dans le calcul approché des valeurs de la fonction donnée $f(x)$ pour les valeurs de l'argument x qui diffèrent de celles des points d'interpolation. Cette opération s'appelle *interpolation de la fonction* $f(x)$. On distingue une *interpolation au sens strict* lorsque $x \in [x_0, x_n]$, c'est-à-dire lorsque la valeur de x est intermédiaire entre x_0 et x_n, et une *extrapolation* lorsque $x \overline{\in} [x_0, x_n]$. Dans ce qui suit, nous entendrons par *interpolation* la première ainsi que la deuxième opération.

§ 5. Première formule d'interpolation de Newton

Soit $y_i = f(x_i)$ les valeurs données de la fonction $y = f(x)$ pour des valeurs équidistantes de la variable indépendante $x_i = x_0 + ih$ $(i = 0, 1, 2, \ldots, n)$, où h est le *pas d'interpolation*. On se propose de choisir un polynôme $P_n(x)$ de degré inférieur ou égal à n, qui prend aux points x_i les valeurs

$$P_n(x_i) = y_i \quad (i = 0, 1, \ldots, n). \tag{1}$$

Les conditions (1) sont équivalentes à ce que

$$\Delta^m P_n(x_0) = \Delta^m y_0$$

pour $m = 0, 1, 2, \ldots, n$.

Recherchons d'après Newton le polynôme sous la forme

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \\ + a_3(x - x_0)(x - x_1)(x - x_2) + \ldots \\ \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \quad (2)$$

La puissance généralisée permet d'écrire pour l'expression (1)

$$P_n(x) = a_0 + a_1(x - x_0)^{[1]} + a_2(x - x_0)^{[2]} + \\ + a_3(x - x_0)^{[3]} + \ldots + a_n(x - x_0)^{[n]}. \quad (2')$$

Notre tâche consiste à déterminer les coefficients a_i ($i = 0, 1, 2, \ldots \ldots, n$) du polynôme $P_n(x)$. Posant dans (2') $x = x_0$, on aura:

$$P_n(x_0) = y_0 = a_0.$$

Pour trouver le coefficient a_1 composons la différence première

$$\Delta P_n(x) = a_1 h + 2a_2(x - x_0)^{[1]} h + \\ + 3a_3(x - x_0)^{[2]} h + \ldots + na_n(x - x_0)^{[n-1]} h.$$

Supposant que dans cette dernière expression $x = x_0$, on obtient:

$$\Delta P_n(x_0) = \Delta y_0 = a_1 h,$$

d'où

$$a_1 = \frac{\Delta y_0}{1!\,h}.$$

Pour déterminer le coefficient a_2 composons la différence seconde

$$\Delta^2 P_n(x) = 2!\,h^2 a_2 + 2 \cdot 3h^2 a_3(x - x_0)^{[1]} + \ldots + \\ + (n - 1)\,nh^2 a_n(x - x_0)^{[n-2]}.$$

Posant $x = x_0$, on obtient:

$$\Delta^2 P_n(x_0) = \Delta^2 y_0 = 2!\,h^2 a_2;$$

d'où

$$a_2 = \frac{\Delta^2 y_0}{2!\,h^2}.$$

De proche en proche, on trouve que

$$a_i = \frac{\Delta^i y_0}{i!\,h^i} \quad (i = 0, 1, 2, \ldots n),$$

où l'on a posé

$$0! = 1 \quad \text{et} \quad \Delta^0 y = y.$$

Portant les valeurs obtenues des coefficients a_i dans l'expression (2') on aboutit au *polynôme d'interpolation de Newton*

$$P_n(x) = y_0 + \frac{\Delta y_0}{1!\,h}(x-x_0)^{[1]} + \frac{\Delta^2 y_0}{2!\,h^2}(x-x_0)^{[2]} + \ldots + \frac{\Delta^n y_0}{n!\,h^n}(x-x_0)^{[n]}. \quad (3)$$

On voit facilement que le polynôme (3) vérifie parfaitement les restrictions du problème posé. En effet, premièrement le degré du polynôme $P_n(x)$ est égal ou inférieur à n, deuxièmement

$$P_n(x_0) = y_0$$

et

$$P_n(x_k) = y_0 + \frac{\Delta y_0}{h}(x_k - x_0) + \frac{\Delta^2 y_0}{2!\,h^2}(x_k - x_0)(x_k - x_1) + \ldots + \frac{\Delta^k y_0}{k!\,h^k}(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1}) =$$
$$= y_0 + k\Delta y_0 + \frac{k(k-1)}{2!}\Delta^2 y_0 + \ldots + \frac{k(k-1)\ldots 1}{k!}\Delta^k y_0 =$$
$$= (1+\Delta)^k y_0 = y_k \quad (k = 1, 2, \ldots, n).$$

Notons que pour $h \to 0$ la formule (3) se ramène au polynôme de Taylor pour la fonction y.

En effet,

$$\lim_{h\to 0} \frac{\Delta^k y_0}{h^k} = \left(\frac{d^k y}{dx^k}\right)_{x=x_0} = y^{(k)}(x_0).$$

De plus, il est évident que

$$\lim_{h\to 0}(x-x_0)^{[n]} = (x-x_0)^n.$$

Il en résulte que quand $h \to 0$, la formule (3) prend la forme du polynôme de Taylor:

$$P_n(x) = y(x_0) + y'(x_0)(x-x_0) + \ldots + \frac{y^{(n)}(x_0)}{n!}(x-x_0)^n.$$

Pour rendre plus commode l'utilisation pratique de la formule de Newton (3) on l'écrit sous une forme quelque peu différente. A cette fin introduisons une nouvelle variable d'après la formule

$$q = \frac{x-x_0}{h};$$

alors il vient

$$\frac{(x-x_0)^{[i]}}{h^i} = \frac{(x-x_0)}{h}\cdot\frac{(x-x_0-h)}{h}\cdot\frac{(x-x_0-2h)}{h}\cdots$$
$$\cdots\frac{[x-x_0-(i-1)h]}{h} = q(q-1)(q-2)\ldots(q-i+1)$$
$$(i = 1, 2, \ldots, n).$$

En portant ces expressions dans la formule (3), on aura:

$$P_n(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots + \\ + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0, \quad (4)$$

où $q = \frac{x-x_0}{h}$ est le *nombre de pas* nécessaire pour atteindre le point x en partant du point x_0. C'est la forme définitive de la *première formule d'interpolation de Newton.*

La formule (4) présente un avantage lorsque la fonction $y = f(x)$ est interpolée dans le voisinage de la valeur initiale x_0, où q est petit en valeur absolue.

Si dans (4) on pose $n = 1$, on obtient la formule *d'interpolation linéaire*

$$P_1(x) = y_0 + q\Delta y_0.$$

Avec $n = 2$ on a la formule *d'interpolation parabolique ou quadratique*

$$P_2(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2}\Delta^2 y_0.$$

Si le tableau donné des valeurs de la fonction y est infini, le nombre n dans la formule d'interpolation (4) peut être quelconque. Pratiquement dans ce cas on le choisit tel que la différence $\Delta^n y_i$ soit constante avec la précision imposée. Pour valeur initiale x_0 on peut prendre toute valeur tabulée de l'argument x.

Si le tableau des valeurs de la fonction est fini, le nombre n est borné et, notamment, n ne peut être supérieur au nombre de valeurs de la fonction y diminué d'une unité.

Remarquons qu'en appliquant la première formule d'interpolation de Newton (par différences descendantes), il est commode de recourir au tableau horizontal, les valeurs nécessaires des différences de la fonction figurant sur la ligne horizontale correspondante du tableau.

Exemple 1. Le pas étant $h = 0{,}05$, construire sur le segment [3,5; 3,6] le polynôme d'interpolation de Newton pour la fonction $y = e^x$ donnée par le tableau

x	3,50	3,55	3,60	3,65	3,70
y	33,115	34,813	36,598	38,475	40,447

Solution. Formons le tableau des différences (tableau 38). Remarquons que, comme d'ordinaire, dans les colonnes des différences nous n'indiquons pas la place de la virgule, qu'on trouve dans la colonne des valeurs de la fonction. Les différences troisièmes étant

Tableau 38

Différences de la fonction $y = e^x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
3,50	33,115	1698	87	5
3,55	34,813	1785	92	3
3,60	36,598	1877	95	
3,65	38,475	1972		
3,70	40,447			

pratiquement constantes, on pose dans la formule (4) $n = 3$. Adoptant $x_0 = 3{,}50$, $y_0 = 33{,}115$, on aura:

$$P_3(x) = 33{,}115 + 1{,}698q + 0{,}087\,\frac{q(q-1)}{2} + 0{,}005\,\frac{q(q-1)(q-2)}{6}$$

ou

$$P_3(x) = 33{,}115 + 1{,}698q + 0{,}0435q(q-1) + 0{,}00083q(q-1)(q-2),$$

avec

$$q = \frac{x-3{,}50}{0{,}50} = 20(x-3{,}5).$$

E x e m p l e 2. Le tableau 39 donne les valeurs de l'intégrale de probabilité

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2}\,dx.$$

En appliquant la première formule d'interpolation de Newton, trouver la valeur approchée de Φ (1,43).

S o l u t i o n. Complétons le tableau 39 de y jusqu'aux différences troisièmes y comprises.

Tableau 39

Différences de la fonction $y = \Phi(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1,0	0,8427	375	−74	10
1,1	0,8802	301	−64	10
1,2	0,9103	237	−54	9
1,3	0,9340	183	−45	9
1,4	0,9523	138	−36	9
1,5	0,9661	102	−27	5
1,6	0,9763	75	−22	6
1,7	0,9838	53	−16	4
1,8	0,9891	37	−12	
1,9	0,9928	25		
2,0	0,9953			

Prenons pour x_0 la valeur tabulée la plus proche de la valeur $x = 1,43$, c'est-à-dire posons $x_0 = 1,4$. Comme $h = 0,1$,

$$q = \frac{1,43 - 1,4}{0,1} = 0,3.$$

En portant dans la formule (4), on obtient :

$$y \approx 0,9523 + 0,3 \cdot 0,0138 + \frac{0,3\,(0,3-1)}{2\,!}\,(-0,0036) + \\ + \frac{0,3\,(0,3-1)\,(0,3-2)}{3\,!} \cdot 0,0009 = 0,95686.$$

(Valeur tabulée: $\Phi\,(1,43) = 0,9569$; cf. « Tables des fonctions » de Yanke et Emde.)

Il arrive souvent en pratique qu'il faut choisir une formule analytique traduisant avec une certaine précision les valeurs tabulées données de la fonction considérée. Une telle formule est dite *empirique*; le problème admet plusieurs solutions.

Pour construire une formule empirique il faut prendre en considération les propriétés générales de la fonction. Si le tableau des différences révèle que les différences d'ordre n de la fonction pour des valeurs équidistantes de l'argument sont constantes, on peut prendre comme formule empirique la première formule d'interpolation de Newton correspondante.

E x e m p l e 3. Construire une formule empirique de la fonction y donnée par le tableau

x	0	1	2	3	4	5
y	5,2	8,0	10,4	12,4	14,0	15,2

S o l u t i o n. Le tableau des différences (tableau 40) montre que la différence seconde est constante. En utilisant la formule d'interpolation de Newton sous la forme (3) et compte tenu de ce que $h = 1$, on aura:

$$y = 5,2 + 2,8x - \frac{0,4}{2}\,x\,(x-1)$$

ou

$$y = 5,2 + 3x - 0,2x^2.$$

E x e m p l e 4. Trouver la somme des carrés

$$S_n = 1^2 + 2^2 + \ldots + n^2$$

des nombres naturels de 1 à n.

Solution. On a évidemment:

$$\Delta S_n = S_{n+1} - S_n = (n + 1)^2.$$

D'où

$$\Delta^2 S_n = 2n + 3, \quad \Delta^3 S_n = 2$$

et, par conséquent, S_n peut être recherchée sous forme de polynôme du troisième degré par rapport à n.

Tableau 40

Différences de la fonction y

x	y	Δy	$\Delta^2 y$
0	5,2	2,8	−0,4
1	8,0	2,4	−0,4
2	10,4	2,0	−0,4
3	12,4	1,6	−0,4
4	14,0	1,2	
5	15,2		

Pour déterminer les différences

$$\Delta S_1, \quad \Delta^2 S_1,$$

il faut calculer trois valeurs S_1, S_2 et S_3. On a:

$$S_1 = 1,$$
$$S_2 = S_1 + 2^2 = 1 + 4 = 5,$$
$$S_3 = S_2 + 3^2 = 5 + 9 = 14.$$

D'où

$$\Delta S_1 = 5 - 1 = 4,$$
$$\Delta S_2 = 14 - 5 = 9,$$
$$\Delta^2 S_1 = 9 - 4 = 5,$$

et

$$\Delta^3 S_1 = 2.$$

En appliquant la première formule de Newton et en tenant compte de ce que

$$q = \frac{n-1}{1} = n - 1,$$

on a:

$$S_n = 1 + 4(n-1) + \frac{5(n-1)(n-2)}{2} + \frac{2(n-1)(n-2)(n-3)}{6}$$

ou

$$S_n = \frac{1}{6} n(n+1)(2n+1).$$

§ 6. Deuxième formule d'interpolation de Newton

La première formule de Newton est pratiquement incommode pour l'interpolation de la fonction dans la partie finale du tableau. Dans ce cas on recourt à la *deuxième formule d'interpolation* que nous déduisons ci-dessous.

Soit un système des valeurs de la fonction

$$y_i = y(x_i) \quad (i = 0, 1, 2, \ldots, n)$$

pour des valeurs équidistantes de l'argument

$$x_i = x_0 + ih.$$

Construisons le polynôme d'interpolation de la forme suivante:

$$P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \\ + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots \\ \ldots + a_n(x - x_n)(x - x_{n-1}) \ldots (x - x_1),$$

ou, en appliquant la puissance généralisée, on obtient:

$$P_n(x) = a_0 + a_1(x - x_n)^{[1]} + a_2(x - x_{n-1})^{[2]} + \\ + a_3(x - x_{n-2})^{[3]} + \ldots + a_n(x - x_1)^{[n]}. \quad (1)$$

Notre tâche consiste à calculer les coefficients $a_0, a_1, a_2, a_3, \ldots, a_n$ de sorte que les égalités

$$P_n(x_i) = y_i \quad (i = 0, 1, 2, \ldots, n)$$

soient vérifiées. A cette fin il faut et il suffit que

$$\Delta^i P_n(x_{n-i}) = \Delta^i y_{n-i} \quad (i = 0, 1, \ldots, n). \quad (2)$$

Posons dans (1) $x = x_n$. Alors, on aura:

$$P_n(x_n) = y_n = a_0,$$

et donc

$$a_0 = y_n.$$

Prenons ensuite les différences premières des premier et deuxième membres de (1)

$$\Delta P_n(x) = a_1 \cdot 1h + a_2 \cdot 2h(x - x_{n-1})^{[1]} + \\ + a_3 \cdot 3h(x - x_{n-2})^{[2]} + \ldots + a_n nh(x - x_1)^{[n-1]}.$$

On en tire en posant $x = x_{n-1}$ et compte tenu des relations (2):

$$\Delta P_n(x_{n-1}) = \Delta y_{n-1} = a_1 h.$$

Par suite

$$a_1 = \frac{\Delta y_{n-1}}{h}.$$

Formant de même la différence seconde de $P_n(x)$ on obtient:

$$\Delta^2 P_n(x) = a_2 2!\, h^2 + a_3 3\cdot 2h^2 (x - x_{n-2})^{[1]} + \ldots + a_n n(n-1)\, h^2 (x - x_1)^{[n-2]}.$$

Posant $x = x_{n-2}$, on trouve:

$$\Delta^2 P_n(x_{n-2}) = \Delta^2 y_{n-2} = a_2 2!\, h^2,$$

ainsi donc

$$a_2 = \frac{\Delta^2 y_{n-2}}{2!\, h^2}.$$

La loi qui régit les coefficients a_i est suffisamment claire. On peut donner une démonstration rigoureuse par récurrence du fait que

$$a_i = \frac{\Delta^i y_{n-i}}{i!\, h^i} \quad (i = 0, 1, 2, \ldots, n). \tag{3}$$

En substituant ces valeurs dans la formule (1) on a finalement:

$$P_n(x) = y_n + \frac{\Delta y_{n-1}}{1!\, h}(x - x_n) + \frac{\Delta^2 y_{n-2}}{2!\, h^2}(x - x_n)(x - x_{n-1}) + \frac{\Delta^3 y_{n-3}}{3!\, h^3}(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots + \frac{\Delta^n y_0}{n!\, h^n}(x - x_n) \ldots (x - x_1). \tag{4}$$

La formule (4) s'appelle *deuxième formule d'interpolation de Newton.*

Introduisons une écriture plus commode de la formule (4). Soit

$$q = \frac{x - x_n}{h},$$

alors

$$\frac{x - x_{n-1}}{h} = \frac{x - x_n + h}{h} = q + 1,$$

$$\frac{x - x_{n-2}}{h} = q + 2, \text{ etc.}$$

En portant ces valeurs dans la formule (4), on obtient:

$$P_n(x) = y_n + q\Delta y_{n-1} + \frac{q(q+1)}{2!}\Delta^2 y_{n-2} + \frac{q(q+1)(q+2)}{3!}\Delta^3 y_{n-3} + \ldots + \frac{q(q+1)\ldots(q+n-1)}{n!}\Delta^n y_0. \tag{4'}$$

C'est précisément la forme usuelle de la *deuxième formule de Newton.* Pour le calcul approché des valeurs de la fonction y on pose:

$$y = P_n(x).$$

E x e m p l e 1. Soit le tableau des valeurs $y = \lg x$ des logarithmes à sept décimales

x	y
1000	3,0000000
1010	3,0043214
1020	3,0086002
1030	3,0128372
1040	3,0170333
1050	3,0211893

Trouver lg 1044.

S o l u t i o n. Formons le tableau des différences (tableau 41).

Tableau 41

Différences de la fonction $y = \lg x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1000	3,0000000	43214	—426	8
1010	3,0043214	42788	—418	9
1020	3,0086002	42370	—409	8
1030	3,0128372	41961	—401	
1040	3,0170333	41560		
1050	3,0211893			

Adoptons

$$x_n = 1050,$$

alors

$$q = \frac{x - x_n}{h} = \frac{1044 - 1050}{10} = -0,6.$$

En utilisant les différences soulignées, on a, en vertu de la formule (4′):

$$\lg 1044 = 3,0211893 + (-0,6) \cdot 0,0041560 + \frac{(-0,6)\cdot(-0,6+1)}{2} \times$$

$$\times 0,0000401 + \frac{(-0,6)\cdot(-0,6+1)\cdot(-0,6+2)}{6} \cdot 0,0000008 = 3,0187005.$$

Dans le résultat obtenu tous les chiffres sont exacts.

Les deux formules de Newton peuvent être utilisées pour extrapoler la fonction, c'est-à-dire pour calculer les valeurs de y pour des valeurs de x dépassant les limites du tableau. Si $x < x_0$ et x est

proche de x_0, la première formule présente plus d'avantages. Dans ce cas

$$q = \frac{x - x_0}{h} < 0.$$

Si $x > x_n$ et x est proche de x_n, il est plus commode de faire appel à la deuxième formule de Newton, et on a

$$q = \frac{x - x_n}{h} > 0.$$

Ainsi la première formule de Newton est appliquée généralement pour *interpoler en avant* et *extrapoler en arrière*, alors que la deuxième s'emploie à l'inverse pour *interpoler en arrière* et *extrapoler en avant.*

Remarquons que dans le cas général l'extrapolation est une opération moins précise qu'une interpolation au sens strict.

E x e m p l e 2. Trouver sin 14° et sin 56° à partir du tableau des valeurs de la fonction $y = \sin x$ entre 15° et 55°, le pas étant $h = 5°$ (tableau 42).

Tableau 42

Différences de la fonction $y = \sin x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
15°	0,2588	832	−26	−6
20°	0,3420	806	−32	−6
25°	0,4226	774	−38	−6
30°	0,5000	736	−44	−5
35°	0,5736	692	−49	−5
40°	0,6428	643	−54	−3
45°	0,7071	589	−57	
50°	0,7660	532		
55°	0,8192			

S o l u t i o n. Formons le tableau des différences (tableau 42). On voit que les différences troisièmes de y sont pratiquement constantes et nous pouvons donc en rester là.

Pour trouver sin 14° posons:

$$x_0 = 15° \quad \text{et} \quad x = 14°;$$

d'où

$$q = \frac{14° - 15°}{5°} = -0,2$$

En appliquant la première formule de Newton et en opérant avec les différences soulignées, on aura:

$$\sin 14° = 0,2588 + (-0,2)\cdot 0,0832 + \frac{(-0,2)(-1,2)}{2\,!}(-0,0026) +$$
$$+ \frac{(-0,2)(-1,2)(-2,2)}{3\,!}(-0,0006) = 0,2419.$$

D'après la table sin 14° = 0,24192.

Pour calculer sin 56° posons:

$$x_n = 55° \quad \text{et} \quad x = 56°;$$

d'où

$$q = \frac{56° - 55°}{5°} = 0,2.$$

En appliquant la deuxième formule de Newton et en utilisant les différences soulignées de deux traits, on aura:

$$\sin 56° = 0,8192 + 0,2\cdot 0,0532 + \frac{0,2\cdot 1,2}{2\,!}(-0,0057) +$$
$$+ \frac{0,2\cdot 1,2\cdot 2,2}{3\,!}(-0,0003) = 0,8291.$$

La table donne sin 56° = 0,82904.

§ 7. Tableau des différences centrales

Pour construire les formules de Newton on ne recourt qu'aux valeurs situées d'un seul côté de la valeur initiale choisie, ce qui donne à ces formules un caractère unilatéral.

Dans de nombreux cas on a intérêt à mettre en œuvre des formules d'interpolation concernant les valeurs situées des deux côtés du point de départ. Les plus usitées sont celles qui contiennent les différences données par la ligne horizontale du tableau diagonal correspondant aux valeurs initiales x_0 et y_0, ou par les lignes immédiatement adjacentes. Ces différences Δy_{-1}, Δy_0, $\Delta^2 y_{-1}$, ... sont dites *différences centrales* (tableau 43), où

$$x_i = x_0 + ih \quad (i = 0, \pm 1, \pm 2, \ldots), \quad y_i = f(x_i),$$
$$\Delta y_i = y_{i+1} - y_i; \quad \Delta^2 y_i = \Delta y_{i+1} - \Delta y_i, \quad \text{etc.}$$

Les formules correspondantes s'appellent *formules d'interpolation par différences centrales*. Ce sont entre autres les formules de Gauss, de Stirling, de Bessel [3].

Tableau 43

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$	$\Delta^6 y$
x_{-4}	y_{-4}						
		Δy_{-4}					
x_{-3}	y_{-3}		$\Delta^2 y_{-4}$				
		Δy_{-3}		$\Delta^3 y_{-4}$			
x_{-2}	y_{-2}		$\Delta^2 y_{-3}$		$\Delta^4 y_{-4}$		
		Δy_{-2}		$\Delta^3 y_{-3}$		$\Delta^5 y_{-4}$	
x_{-1}	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$		$\Delta^6 y_{-4}$
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$	
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$
		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$	
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		$\Delta^6 y_{-2}$
		Δy_1		$\Delta^3 y_0$		$\Delta^5 y_{-1}$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$		
		Δy_2		$\Delta^3 y_1$			
x_3	y_3		$\Delta^2 y_2$				
		Δy_3					
x_4	y_4						

§ 8. Formules d'interpolation de Gauss

Déduisons d'abord les formules de Gauss.

Soient $2n + 1$ points équidistants

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_{-1},\ x_0,\ x_1,\ \ldots,\ x_{n-1},\ x_n,$$

où

$$\Delta x_i = x_{i+1} - x_i = h = \text{const} \quad (i = -n,\ -(n-1),\ \ldots,\ n-1),$$

auxquels on connaît la valeur de la fonction $y = f(x)$

$$y_i = f(x_i) \quad (i = 0,\ \pm 1,\ \ldots,\ \pm n).$$

On demande de construire un polynôme $P(x)$ de degré égal ou inférieur à $2n$ et tel que

$$P(x_i) = y_i \text{ pour } i = 0,\ \pm 1,\ \ldots,\ n$$

La dernière condition entraîne

$$\Delta^k P(x_i) = \Delta^k y_i. \tag{1}$$

pour toutes les valeurs respectives de i et de k.

Cherchons ce polynôme sous la forme

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \\ + a_3(x - x_{-1})(x - x_0)(x - x_1) + a_4(x - x_{-1})(x - x_0)(x - x_1) \times \\ \times (x - x_2) + a_5(x - x_{-2})(x - x_{-1})(x - x_0)(x - x_1)(x - x_2) + \\ \ldots + a_{2n-1}(x - x_{-(n-1)}) \ldots (x - x_{-1})(x - x_0)(x - x_1) \ldots \\ \ldots (x - x_{n-1}) + a_{2n}(x - x_{-(n-1)}) \ldots (x - x_{-1})(x - x_0)(x - x_1) \ldots \\ \ldots (x - x_{n-1})(x - x_n). \tag{2}$$

Introduisant les puissances généralisées, on obtient:

$$P(x) = a_0 + a_1(x - x_0)^{[1]} + a_2(x - x_0)^{[2]} + \\ + a_3(x - x_{-1})^{[3]} + a_4(x - x_{-1})^{[4]} + \ldots + a_{2n-1}(x - x_{-(n-1)})^{[2n-1]} + \\ + a_{2n}(x - x_{-(n-1)})^{[2n]}. \tag{3}$$

En appliquant pour calculer les coefficients a_i ($i = 0, 1, \ldots, 2n$) le même procédé que pour déduire les formules d'interpolation de Newton et en tenant compte de la formule (1), on trouve successivement:

$$a_0 = y_0, \quad a_1 = \frac{\Delta y_0}{1!\,h}, \quad a_2 = \frac{\Delta^2 y_{-1}}{2!\,h^2}, \quad a_3 = \frac{\Delta^3 y_{-1}}{3!\,h^3},$$

$$a_4 = \frac{\Delta^4 y_{-2}}{4!\,h^4}, \ \ldots, \ a_{2n-1} = \frac{\Delta^{2n-1} y_{-(n-1)}}{(2n-1)!\,h^{2n-1}}, \quad a_{2n} = \frac{\Delta^{2n} y_{-n}}{(2n)!\,h^{2n}}.$$

Introduisons ensuite la variable

$$q = \frac{x - x_0}{h}$$

et, après avoir effectué la substitution nécessaire dans la formule (3), on obtient la *première formule d'interpolation de Gauss*

$$P(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_{-1} + \frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-1} + \\ + \frac{(q+1)q(q-1)(q-2)}{4!}\Delta^4 y_{-2} + \frac{(q+2)(q+1)q(q-1)(q-2)}{5!}\Delta^5 y_{-2} + \ldots \\ \ldots + \frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1} y_{-(n-1)} + \\ + \frac{(q+n-1)\ldots(q-n)}{(2n)!}\Delta^{2n} y_{n} \tag{4}$$

ou en abrégé

$$P(x) = y_0 + q\Delta y_0 + \frac{q^{[2]}}{2!}\Delta^2 y_{-1} + \frac{(q+1)^{[3]}}{3!}\Delta^3 y_{-1} +$$
$$+ \frac{(q+1)^{[4]}}{4!}\Delta^4 y_{-2} + \ldots + \frac{(q+n-1)^{[2n-1]}}{(2n-1)!}\Delta^{2n-1} y_{-(n-1)} +$$
$$+ \frac{(q+n-1)^{[2n]}}{(2n)!}\Delta^{2n} y_{-n} \qquad (4')$$

avec $x = x_0 + qh$ et $q^{[m]} = q(q-1)\ldots[q-(m-1)]$.

La première formule de Gauss contient les différences centrales

$$\Delta y_0,\ \Delta^2 y_{-1},\ \Delta^3 y_{-1},\ \Delta^4 y_{-2},\ \Delta^5 y_{-2},\ \Delta^6 y_{-3},\ldots$$

(cf. tableau 43, où ces différences forment la ligne brisée inférieure suivant la flèche). D'une façon analogue on peut obtenir la *deuxième formule d'interpolation de Gauss* qui contient les différences centrales

$$\Delta y_{-1},\ \Delta^2 y_{-1},\ \Delta^3 y_{-2},\ \Delta^4 y_{-2},\ \Delta^5 y_{-3},\ \Delta^6 y_{-3},\ldots$$

(dans le tableau 43 ces différences forment la ligne brisée supérieure suivant la flèche).

La deuxième formule de Gauss s'écrit

$$P(x) = y_0 + q\Delta y_{-1} + \frac{(q+1)q}{2!}\Delta^2 y_{-1} + \frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-2} +$$
$$+ \frac{(q+2)(q+1)q(q-1)}{4!}\Delta^4 y_{-2} + \ldots + \frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1} y_{-n} +$$
$$+ \frac{(q+n)(q+n-1)\ldots(q-n+1)}{(2n)!}\Delta^{2n} y_{-n} \qquad (5)$$

ou, en notations abrégées,

$$P(x) = y_0 + q\Delta y_{-1} + \frac{(q+1)^{[2]}}{2!}\Delta^2 y_{-1} +$$
$$+ \frac{(q+1)^{[3]}}{3!}\Delta^3 y_{-2} + \frac{(q+2)^{[4]}}{4!}\Delta^4 y_{-2} + \ldots$$
$$\ldots + \frac{(q+n-1)^{[2n-1]}}{(2n-1)!}\Delta^{2n-1} y_{-n} + \frac{(q+n)^{[2n]}}{(2n)!}\Delta^{2n} y_{-n} \qquad (5')$$

avec

$$x = x_0 + qh.$$

§ 9. Formule d'interpolation de Stirling

Prenant la moyenne arithmétique de la première et de la deuxième formule de Gauss (4) et (5) (§ 8), on obtient la *formule de Stirling*:

$$P(x)=y_0+q\cdot\frac{\Delta y_{-1}+\Delta y_0}{2}+\frac{q^2}{2}\Delta^2 y_{-1}+\frac{q(q^2-1^2)}{3!}\cdot\frac{\Delta^3 y_{-2}+\Delta^3 y_{-1}}{2}+$$
$$+\frac{q^2(q^2-1^2)}{4!}\Delta^4 y_{-2}+\frac{q(q^2-1^2)(q^2-2^2)}{5!}\cdot\frac{\Delta^5 y_{-3}+\Delta^5 y_{-2}}{2}+$$
$$+\frac{q^2(q^2-1^2)(q^2-2^2)}{6!}\Delta^6 y_{-3}+\ldots+$$
$$+\frac{q(q^2-1^2)(q^2-2^2)(q^2-3^2)\ldots[q^2-(n-1)^2]}{(2n-1)!}\times$$
$$\times\frac{\Delta^{2n-1}y_{-n}+\Delta^{2n-1}y_{-(n-1)}}{2}+\frac{q^2(q^2-1^2)(q^2-2^2)\ldots[q^2-(n-1)^2]}{(2n)!}\Delta^{2n}y_{-n},$$

où

$$q=\frac{x-x_0}{h}.$$

On voit aisément que

$$P(x_i)=y_i \quad \text{pour} \quad i=0,\ \pm 1,\ \ldots,\ \pm n.$$

§ 10. Formule d'interpolation de Bessel

Outre la *formule de Stirling*, on utilise souvent la *formule de Bessel*. Pour la déduire faisons appel à la deuxième formule de Gauss (5) (cf. § 8).

Prenons $2n+1$ points d'interpolation équidistants

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_0,\ \ldots,\ x_{n-1},\ x_n,$$

au pas h et soit

$$y_i=f(x_i) \quad (i=-n,\ \ldots,\ n+1)$$

les valeurs données de la fonction $y=f(x)$.

Si l'on prend pour valeurs initiales $x=x_0$ et $y=y_0$, en utilisant les points x_k $(k=0,\ \pm 1,\ \ldots,\ \pm n)$ on a:

$$P(x)=y_0+q\Delta y_{-1}+\frac{(q+1)q}{2!}\Delta^2 y_{-1}+$$
$$+\frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-2}+\frac{(q+2)(q+1)q(q-1)}{4!}\Delta^4 y_{-2}+\ldots$$
$$\ldots+\frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1}y_{-n}+$$
$$+\frac{(q+n)(q+n-1)\cdots(q-n+1)}{(2n)!}\Delta^{2n}y_{-n}. \quad (1)$$

Prenons maintenant comme valeurs initiales $x = x_1$ et $y = y_1$ et utilisons les points x_{1+k} $(k = 0, \pm 1, \ldots, \pm n)$. Il vient

$$\frac{x-x_1}{h} = \frac{x-x_0-h}{h} = q-1,$$

les indices de toutes les différences du deuxième membre de (1) augmentant respectivement de l'unité. Si dans le deuxième membre de (1) on remplace q par $q-1$ tout en augmentant de l'unité les indices de toutes les différences, on obtient la formule auxiliaire

$$P(x) = y_1 + (q-1)\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \\ + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_{-1} + \frac{(q+1)q(q-1)(q-2)}{4!}\Delta^4 y_{-1} + \\ + \frac{(q+1)q(q-1)(q-2)(q-3)}{5!}\Delta^5 y_{-2} + \ldots + \frac{(q+n-2)\ldots(q-n)}{(2n-1)!} \times \\ \times \Delta^{2n-1} y_{-(n-1)} + \frac{(q+n-1)\ldots(q-n)}{(2n)!}\Delta^{2n} y_{-(n-1)}. \quad (2)$$

En prenant la moyenne arithmétique des formules (4) du § 8 et (2) après des transformations élémentaires on obtient la *formule d'interpolation de Bessel*

$$P(x) = \frac{y_0+y_1}{2} + \left(q-\frac{1}{2}\right)\Delta y_0 + \frac{q(q-1)}{2}\cdot\frac{\Delta^2 y_{-1}+\Delta^2 y_0}{2} + \\ + \frac{\left(q-\frac{1}{2}\right)q(q-1)}{3!}\Delta^3 y_{-1} + \frac{q(q-1)(q+1)(q-2)}{4!}\cdot\frac{\Delta^4 y_{-2}+\Delta^4 y_{-1}}{2} + \\ + \frac{\left(q-\frac{1}{2}\right)q(q-1)(q+1)(q-2)}{5!}\Delta^5 y_{-2} + \\ + \frac{q(q-1)(q+1)(q-2)(q+2)(q-3)}{6!}\cdot\frac{\Delta^6 y_{-3}+\Delta^6 y_{-2}}{2} + \ldots \\ \ldots + \frac{q(q-1)(q+1)(q-2)(q+2)\ldots(q-n)(q+n-1)}{(2n)!} \times \\ \times \frac{\Delta^{2n} y_{-n}+\Delta^{2n} y_{-n+1}}{2} + \\ + \frac{\left(q-\frac{1}{2}\right)q(q-1)(q+1)(q-2)(q+2)\ldots(q-n)(q+n-1)}{(2n+1)!}\Delta^{2n+1} y_{-n}, \quad (3)$$

où

$$q = \frac{x-x_0}{h}.$$

Comme le montre la déduction, la formule de Bessel (3) est un polynôme qui coïncide avec la fonction donnée $y = f(x)$ en $2n+2$

points

$$x_{-n},\ x_{-(n-1)},\ \ldots,\ x_n,\ x_{n+1}.$$

Dans le cas particulier, pour $n = 1$, en négligeant la différence $\Delta^3 y_{-1}$ on a la *formule d'interpolation quadratique de Bessel*

$$P(x) = \frac{y_0 + y_0 + \Delta y_0}{2} + \left(q - \frac{1}{2}\right)\Delta y_0 + \frac{q(q-1)}{2}\cdot\frac{\Delta y_0 - \Delta y_{-1} + \Delta y_1 - \Delta y_0}{2}$$

ou

$$P(x) = y_0 + q\Delta y_0 - q_1(\Delta y_1 - \Delta y_{-1})$$

avec

$$q_1 = \frac{q(1-q)}{4}.$$

Dans la formule de Bessel tous les termes contenant les différences d'ordre impair comportent le facteur $q - \frac{1}{2}$; c'est pourquoi avec $q = \frac{1}{2}$ la formule (3) devient beaucoup plus simple:

$$P\left(\frac{x_0 + x_1}{2}\right) = \frac{y_0 + y_1}{2} - \frac{1}{8}\,\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{3}{128}\,\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} - \frac{5}{1024}\,\frac{\Delta^6 y_{-3} + \Delta^6 y_{-2}}{2} + \ldots + (-1)^n \frac{[1\cdot 3\cdot 5 \ldots (2n-1)]^2}{2^{2n}(2n)!}\,\frac{\Delta^{2n} y_{-n} + \Delta^{2n} y_{-n+1}}{2}.$$

Ce cas spécial s'appelle *formule de dichotomie de Bessel.* Si dans la formule (3) on effectue le changement de variable d'après la formule $q - \frac{1}{2} = p$, la formule devient plus symétrique:

$$\begin{aligned} P(x) ={}& \frac{y_0 + y_1}{2!} + p\Delta y_0 + \frac{\left(p^2 - \frac{1}{4}\right)}{2}\cdot\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \\ &+ \frac{p\left(p^2 - \frac{1}{4}\right)}{3!}\Delta^3 y_{-1} + \frac{\left(p^2 - \frac{1}{4}\right)\left(p^2 - \frac{9}{4}\right)}{4!}\cdot\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \\ &+ \frac{p\left(p^2 - \frac{1}{4}\right)\left(p^2 - \frac{9}{4}\right)}{5!}\Delta^5 y_{-2} + \frac{\left(p^2 - \frac{1}{4}\right)\left(p^2 - \frac{9}{4}\right)\left(p^2 - \frac{25}{4}\right)}{6!} \times \\ &\times \frac{\Delta^6 y_{-3} + \Delta^6 y_{-2}}{2} + \ldots + \frac{\left(p^2 - \frac{1}{4}\right)\left(p^2 - \frac{9}{4}\right)\ldots\left[p^2 - \frac{(2n-1)^2}{4}\right]}{(2n)!} \times \\ &\times \frac{\Delta^{2n} y_{-n} + \Delta^{2n} y_{-n+1}}{2} + \frac{p\left(p^2 - \frac{1}{4}\right)\left(p^2 - \frac{9}{4}\right)\ldots\left[p^2 - \frac{(2n-1)^2}{4}\right]}{(2n+1)!} \times \\ &\times \Delta^{2n+1} y_{-n+1}, \end{aligned} \qquad (3')$$

où $p = \frac{1}{h}\left(x - \frac{x_0 + x_1}{2}\right)$.

§ 11. Caractéristique générale des formules d'interpolation à pas constant

Pour donner une caractéristique générale des formules d'interpolation notons que dans le cas des formules de Newton on prend comme valeur initiale x_0 le premier et le dernier point d'interpolation ; dans le cas de l'interpolation par différences centrales, le point initial est médian. Le tableau 44 schématise l'utilisation des différences dans les formules d'interpolation principales. L'ordre d'indexage dans la deuxième formule de Newton est changé pour rendre plus commode la lecture du tableau.

Tableau 44

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	Notes
						2e formule de Newton
x_{-2}	y_{-2}		$\Delta^2 y_{-3}$		$\Delta^4 y_{-4}$	
		Δy_{-2}		$\Delta^3 y_{-3}$		
x_{-1}	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$	
		Δy_{-1}		$\Delta^3 y_{-2}$		
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$	Formule de Stirling
		Δy_0		$\Delta^3 y_{-1}$		Formule de Bessel
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$	
		Δy_1		$\Delta^3 y_0$		
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$	
		Δy_2		$\Delta^3 y_1$		1re formule de Newton
x_3	y_3		$\Delta^2 y_2$		$\Delta^4 y_1$	

Une étude plus poussée des formules montre que pour $|q| \leqslant 0{,}25$ il vaut mieux appliquer la formule de Stirling, et pour $0{,}25 \leqslant q \leqslant 0{,}75$, celle de Bessel. Il est avantageux d'appliquer la première et la deuxième formule de Newton lorsque l'interpolation porte sur le début ou respectivement sur la fin du tableau et les différences centrales nécessaires font défaut [4].

Exemple 1. Les valeurs de l'intégrale de probabilité [3]

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2}\, dx$$

sont données dans le tableau 45. Trouver Φ (0,5437).

Tableau 45

Différences de la fonction $y = \Phi(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0,51	0,5292437			
		86550		
0,52	0,5378987		−896	
		85654		−7
0,53	0,5464641		−903	
		84751		−7
0,54	0,5549392		−910	
		83841		−7
0,55	0,5633233		−917	
		82924		−6
0,56	0,5716157		−923	
		82001		
0,57	0,5798158			

S o l u t i o n. Complétons le tableau 45 par les différences finies de la fonction donnée $y = \Phi(x)$. Posons $x_0 = 0{,}54$ et $x = 0{,}5437$; il vient

$$q = \frac{x - x_0}{h} = \frac{0{,}5437 - 0{,}54}{0{,}01} = 0{,}37.$$

Comme $\frac{1}{4} < q < \frac{3}{4}$, appliquons la formule de Bessel (3′). On a :

$$p = q - \frac{1}{2} = 0{,}37 - 0{,}50 = -0{,}13\,;$$

d'où, en utilisant les différences soulignées,

$$\Phi(0{,}5437) = \frac{0{,}5549392 + 0{,}5633233}{2} +$$

$$+ (-0{,}13)\,0{,}0083841 + \frac{0{,}0169 - 0{,}25}{2}\;\frac{-0{,}0000910 - 0{,}0000917}{2} +$$

$$+ \frac{-0{,}13\,(0{,}0169 - 0{,}25)}{6}\,(-0{,}0000007) =$$

$$= 0{,}55913125 - 0{,}00108993 + 0{,}00001065 = 0{,}5580520.$$

E x e m p l e 2. Soit le tableau 46 des valeurs de l'intégrale elliptique totale

$$K(\alpha) = \int_0^{\frac{\pi}{2}} \frac{dx}{\sqrt{1 - \sin^2\alpha \sin^2 x}}\,.$$

Trouver $K\,(78°30')$.

Tableau 46

Valeurs de l'intégrale elliptique totale K (α)

α	$K(\alpha)$	ΔK	$\Delta^2 K$	$\Delta^3 K$	$\Delta^4 K$	$\Delta^5 K$	$\Delta^6 K$
75°	2,76806						
		6461					
76°	2,83267		528				
		6989		84			
77°	2,90256		612		19		
		7601		103		13	
78°	2,97857		715		32		—5
		8316		135		8	
79°	3,06173		850		40		18
		9166		175		26	
80°	3,15339		1025		66		—1
		10191		241		25	
81°	3,25530		1266		91		43
		11457		332		68	
82°	3,36987		1598		159		
		13055		491			
83°	3,50042		2089				
		15144					
84°	3,65186						

S o l u t i o n. Posons $x_0 = 78°$; $h = 1°$; $x = 78°30'$; d'où $q = 0,5$. Si l'on recourt à la formule de dichotomie de Bessel, on aura en se bornant aux différences cinquièmes:

$$K(78°30') = 2,97857 + 0,5 \cdot 8316 \cdot 10^{-5} - 0,125 \cdot \frac{715+850}{2} \cdot 10^{-5} +$$
$$+ 0,023437 \cdot \frac{32+40}{2} \cdot 10^{-5} = 2\ 97857 - 0,04158 -$$
$$- 0\ 000978 + 0,000008 = 3,019180.$$

Pour comparer appliquons maintenant la formule de Stirling

$$K(78°30') = 2,97857 + 0,5\, \frac{7601+8316}{2} \cdot 10^{-5} +$$
$$+ 0,125715 \cdot 10^{-5} - 0\ 0625 \cdot \frac{103+135}{2} \cdot 10^{-5} -$$
$$- 0,0078 \cdot 32 \cdot 10^{-5} + 0\ 0117 \cdot \frac{13+8}{2} \cdot 10^{-5} =$$
$$= 2,97857 + 0,039792 + 0\ 000894 - 0\ 000074 -$$
$$- 0,000002 + 0,000001 = 3,019181.$$

§ 12. Formule d'interpolation de Lagrange

Les formules d'interpolation déduites dans les paragraphes précédents ne sont applicables que dans le cas des points équidistants. Pour des points arbitraires on utilise une formule plus générale appelée *formule d'interpolation de Lagrange.*

Supposons que pour $n+1$ valeurs distinctes de l'argument $x_0, x_1, x_2, \ldots, x_n$ données sur le segment $[a, b]$ on connaisse les valeurs correspondantes de la fonction $y = f(x)$

$$f(x_0) = y_0,$$
$$f(x_1) = y_1, \ldots, f(x_n) = y_n.$$

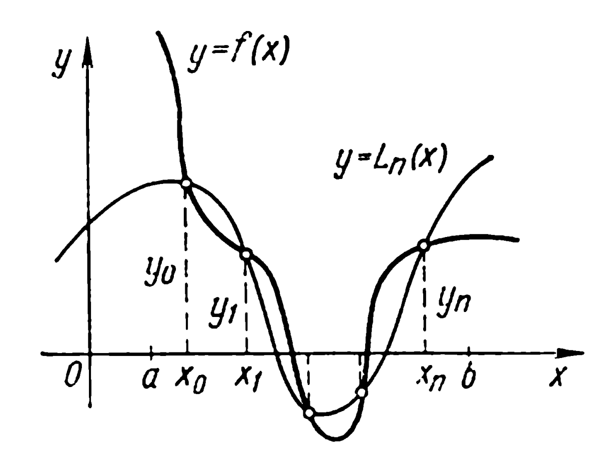


Fig. 62*a*.

On demande de construire le polynôme $L_n(x)$ de degré égal ou inférieur à n et prenant aux points donnés $x_0, x_1, \ldots, x_n$ les mêmes valeurs que la fonction $f(x)$, c'est-à-dire tel que

$$L_n(x_i) = y_i$$
$$(i = 0, 1, 2, \ldots, n)$$

(fig. 62*a*).

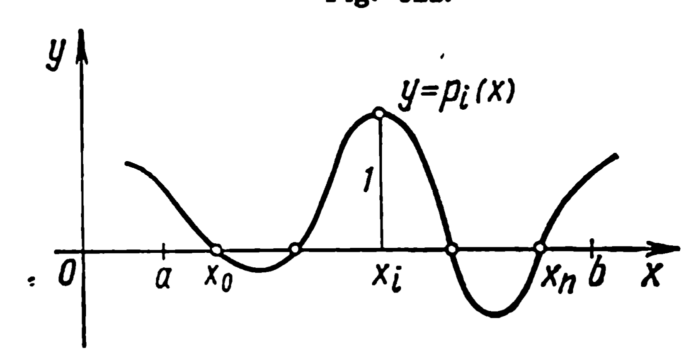


Fig. 62*b*.

Résolvons d'abord le problème partiel : construire un polynôme $p_i(x)$ tel que

$$p_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{si } j = i; \\ 0 & \text{si } j \neq i, \end{cases} \qquad (1)$$

où δ_{ij} est le *symbole de Kronecker* (fig. 62*b*).

Le polynôme à obtenir s'annulant en n points $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, il s'écrit

$$p_i(x) = C_i(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n), \qquad (2)$$

où C_i est une constante. Posant dans la formule (2) $x = x_i$ et prenant en considération que $p_i(x_i) = 1$, on obtient :

$$C_i(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n) = 1.$$

D'où

$$C_i = \frac{1}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}.$$

Portant cette valeur dans la formule (2) on obtient:

$$p_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}. \quad (3)$$

Maintenant passons à la résolution du problème général qui consiste à former le polynôme $L_n(x)$ vérifiant les conditions indiquées ci-dessus: $L_n(x_i) = y_i$.

Ce polynôme est de la forme:

$$L_n(x) = \sum_{i=0}^{n} p_i(x)\, y_i. \quad (4)$$

En effet, premièrement, il est clair que le degré du polynôme construit $L_n(x)$ est égal ou inférieur à n, et deuxièmement, en vertu de la condition (1), on a:

$$L_n(x_j) = \sum_{i=0}^{n} p_i(x_j)\, y_i = p_j(x_j)\, y_j = y_j \quad (j = 0, 1, \ldots, n).$$

Portant dans la formule (4) la valeur de $p_i(x)$ tirée de (3) on obtient l'expression

$$L_n(x) = \sum_{i=0}^{n} y_i \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \quad (5)$$

qui est précisément la *formule d'interpolation de Lagrange.*

Montrons l'*unicité* du polynôme de Lagrange.

Raisonnons par l'absurde.

Soit $\tilde{L}_n(x)$ un polynôme distinct de $L_n(x)$ de degré égal ou inférieur à n et tel que

$$\tilde{L}_n(x_i) = y_i \quad (i = 0, 1, \ldots, n).$$

Alors, le polynôme

$$Q_n(x) = \tilde{L}_n(x) - L_n(x)$$

dont le degré est évidemment égal ou inférieur à n, s'annule en $n+1$ points $x_0, x_1, x_2, \ldots, x_n$, c'est-à-dire

$$Q_n(x) \equiv 0.$$

Par suite,

$$\tilde{L}_n(x) \equiv L_n(x).$$

On en tire en particulier que si les points d'interpolation sont équidistants, le polynôme de Lagrange coïncide avec le polynôme de Newton correspondant.

Remarquons dans le cas général qu'avec un choix de points approprié toutes les formules d'interpolation précédentes se déduisent de la formule de Lagrange.

La formule de Lagrange (5) peut être mise sous une forme plus condensée. A cet effet introduisons la notation

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \qquad (6)$$

En dérivant ce produit par rapport à x on obtient:

$$\Pi'_{n+1}(x) = \sum_{j=0}^{n} (x - x_0)(x - x_1)\ldots(x - x_{j-1})(x - x_{j+1})\ldots(x - x_n).$$

Adoptant $x = x_i$ $(i = 0, 1, 2, \ldots, n)$, on aura:

$$\Pi'_{n+1}(x_i) = (x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_0). \qquad (7)$$

Portant les expressions (6) et (7) dans la formule (5) on obtient:

$$L_n(x) = \Pi_{n+1}(x) \sum_{i=0}^{n} \frac{y_i}{\Pi'_{n+1}(x_i)(x - x_i)}. \qquad (5')$$

Il convient de signaler un fait important: à la différence des formules précédentes, la formule de Lagrange contient y_i sous une forme explicite.

Considérons deux cas particuliers du polynôme de Lagrange.

Pour $n = 1$ nous avons deux points et la formule de Lagrange est dans ce cas l'équation d'une droite $y = L_1(x)$ qui passe par deux points donnés:

$$y = \frac{x - b}{a - b} y_0 + \frac{x - a}{b - a} y_1,$$

où a et b sont les abscisses de ces points.

Pour $n = 2$ on obtient l'équation d'une parabole $y = L_2(x)$ qui passe par trois points:

$$y = \frac{(x - b)(x - c)}{(a - b)(a - c)} y_0 + \frac{(x - a)(x - c)}{(b - a)(b - c)} y_1 + \frac{(x - a)(x - b)}{(c - a)(c - b)} y_2,$$

où a, b, c sont les abscisses des points considérés.

Exemple 1. Construire le polynôme de Lagrange de la fonction $y = \sin \pi x$ pour les points

$$x_0 = 0, \quad x_1 = \frac{1}{6} \quad x_2 = \frac{1}{2}.$$

Solution. Calculons les valeurs correspondantes de la fonction:

$$y_0 = 0, \quad y_1 = \sin\frac{\pi}{6} = \frac{1}{2}, \; y_2 = \sin\frac{\pi}{2} = 1.$$

Appliquant la formule (5), on aura:

$$L_2(x)=\frac{\left(x-\frac{1}{6}\right)\left(x-\frac{1}{2}\right)}{\left(-\frac{1}{6}\right)\left(-\frac{1}{2}\right)}\cdot 0+\frac{x\left(x-\frac{1}{2}\right)}{\frac{1}{6}\left(\frac{1}{6}-\frac{1}{2}\right)}\cdot\frac{1}{2}+\frac{x\left(x-\frac{1}{6}\right)}{\frac{1}{2}\left(\frac{1}{2}-\frac{1}{6}\right)}\cdot 1$$

ou

$$L_2(x)=\frac{7}{2}x-3x^2.$$

Exemple 2. Soit le tableau des valeurs de la fonction $y = f(x)$ [3]:

x	y
321,0	2,50651
322,8	2,50893
324,2	2,51081
325,0	2,51188

Calculer la valeur $f(323{,}5)$.

Solution. Posons $x = 323{,}5$; $n = 3$. Alors d'après la formule (5), on a:

$$\begin{aligned}f(323{,}5)=&\frac{(323{,}5-322{,}8)(323{,}5-324{,}2)(323{,}5-325{,}0)}{(321-322{,}8)(321-324{,}2)(321-325)}\cdot 2{,}50651+\\&+\frac{(323{,}5-321)(323{,}5-324{,}2)(323{,}5-325)}{(322{,}8-321)(322{,}8-324{,}2)(322{,}8-325)}\cdot 2{,}50893+\\&+\frac{(323{,}5-321)(323{,}5-322{,}8)(323{,}5-325)}{(324{,}2-321)(324{,}2-322{,}8)(324{,}2-325)}\cdot 2{,}51081+\\&+\frac{(323{,}5-321)(323{,}5-322{,}8)(323{,}5-324{,}2)}{(325-321)(325-322{,}8)(325-324{,}2)}\cdot 2{,}51188=\\&=-0{,}07996+1{,}18794+1{,}83897-0{,}43708=2{,}50987.\end{aligned}$$

§ 13*. Calcul des coefficients de Lagrange

Indiquons un schéma qui rend plus facile le calcul des coefficients de y_i $(i = 0, 1, 2, \ldots, n)$ dans la formule de Lagrange dits *coefficients de Lagrange*

$$L_i^{(n)}(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}, \quad (1)$$

ou sous une forme abrégée

$$L_i^{(n)}(x)=\frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}, \quad (2)$$

où

$$\Pi_{n+1}(x) = (x - x_0) \ldots (x - x_n).$$

La formule de Lagrange s'écrit alors

$$L_n(x) = \sum_{i=0}^{n} L_i^{(n)}(x)\, y_i.$$

Notons que la forme des coefficients de Lagrange est invariante par rapport à une substitution linéaire entière $x = at + b$ (a, b sont des constantes et $a \neq 0$). En effet, posons dans la formule (1)

$$x = at + b; \quad x_j = at_j + b \quad (j = 0, 1, \ldots, n);$$

en divisant le numérateur et le dénominateur par a^n, on obtient:

$$L_i^{(n)}(t) = \frac{(t-t_0)(t-t_1)\ldots(t-t_{i-1})(t-t_{i+1})\ldots(t-t_n)}{(t_i-t_0)(t_i-t_1)\ldots(t_i-t_{i-1})(t-t_{i+1})\ldots(t_i-t_n)} \qquad (3)$$

ou

$$L_i^{(n)} = \frac{\Pi_{n+1}(t)}{(t-t_i)\,\Pi'_{n+1}(t_i)} \qquad (3')$$

avec

$$\Pi_{n+1}(t) = (t - t_0)(t - t_1) \ldots (t - t_n),$$

ce qu'il fallait démontrer.

Pour calculer les coefficients de Lagrange on peut utiliser le schéma ci-dessous, commode à réaliser sur un calculateur électronique. Rangeons d'abord les différences en un tableau de la façon suivante:

$$\begin{array}{ccccc}
\underline{x - x_0} & x_0 - x_1 & x_0 - x_2 & \ldots & x_0 - x_n \\
x_1 - x_0 & \underline{x - x_1} & x_1 - x_2 & \ldots & x_1 - x_n \\
x_2 - x_0 & x_2 - x_1 & \underline{x - x_2} & \ldots & x_2 - x_n \\
\cdot\cdot\cdot & \cdot\cdot\cdot & \cdot\cdot\cdot & \cdot\cdot\cdot & \cdot\cdot\cdot \\
\cdot\cdot\cdot & \cdot\cdot\cdot & \cdot\cdot\cdot & \cdot\cdot\cdot & \cdot\cdot\cdot \\
x_n - x_0 & x_n - x_1 & x_n - x_2 & \ldots & \underline{x - x_n}.
\end{array} \qquad (*)$$

Désignons le produit des éléments de la première ligne par D_0, de la deuxième ligne par D_1, etc. Quant au produit des éléments de la diagonale principale (éléments du schéma soulignés), il est évident qu'il s'écrit $\Pi_{n+1}(x)$. On en tire que

$$L_i^{(n)}(x) = \frac{\Pi_{n+1}(x)}{D_i} \qquad (i = 0, 1, \ldots, n). \qquad (4)$$

Par conséquent,

$$L_n(x) = \Pi_{n+1}(x) \sum_{i=0}^{n} \frac{y_i}{D_i}. \qquad (5)$$

Dans le cas des points équidistants, les coefficients de Lagrange peuvent être simplifiés.

En effet, en posant

$$x = x_0 + th,$$

on aura:

$$t_0 = 0, \ t_1 = 1, \ \ldots, \ t_n = n.$$

D'où

$$\Pi_{n+1}(t) = t(t-1)(t-2)\ldots(t-n)$$

et

$$\Pi'_{n+1}(i) = (-1)^{n-i} i!\,(n-i)!$$

Portant ces expressions dans la formule (3′), on obtient:

$$L_i^{(n)}(t) = \frac{1}{n!}\Pi_{n+1}(t)\cdot\frac{(-1)^{n-i}C_n^i}{t-i} \quad (i = 0, 1, \ldots n), \tag{6}$$

où

$$C_n^i = \frac{n!}{i!\,(n-i)!}.$$

On en tire

$$L_n(x) = \frac{1}{n!}\Pi_{n+1}(t)\sum_{i=0}^{n}(-1)^{n-i}\frac{C_n^i}{t-i}y_i, \tag{7}$$

où

$$t = \frac{x-x_0}{h}.$$

Dans le cas d'un pas h constant, le problème d'interpolation est rendu encore plus facile par le fait qu'il existe des tables des coefficients de Lagrange (cf. [5]), les calculs se ramenant ainsi à la multiplication des coefficients tabulés par les valeurs correspondantes de la fonction y_i et à la sommation.

Exemple 1. Soit le tableau des valeurs d'une fonction $y = y(x)$

x	0,05	0,15	0,20	0,25	0,35	0,40	0,50	0,55
y	0,9512	0,8607	0,8187	0,7788	0,7047	0,6703	0,6065	0,5769
t	1	3	4	5	7	8	10	11

Trouver $y(0{,}45)$.

Solution. Pour simplifier les calculs, posons:

$$x = 0{,}05t.$$

Alors les valeurs de la nouvelle variable t associées aux points d'interpolation seront 1, 3, 4, 5, 7, 8, 10, 11. Il faut trouver la valeur

de y pour $x = 0{,}45$, c'est-à-dire pour $t = 9$. En adoptant $t = t_i$ ($i = 0, 1, 2, \ldots, 7$), disposons les calculs suivant le schéma ci-dessus (tableau 47).

Tableau 47

Schéma de calcul des coefficients de Lagrange

i	$t_i - t_j$ $(j \neq i)$								D_i	y_i	$\frac{y_i}{D_i}$
0	8	−2	−3	−4	−6	−7	−9	−10	−725 760	0,9512	$-0{,}0131 \cdot 10^{-4}$
1	2	6	−1	−2	−4	−5	−7	−8	26 880	0,8607	$0{,}3202 \cdot 10^{-4}$
2	3	1	5	−1	−3	−4	−6	−7	−7 560	0,8187	$-1{,}0829 \cdot 10^{-4}$
3	4	2	1	4	−2	−3	−5	−6	5 760	0,7788	$1{,}3520 \cdot 10^{-4}$
4	6	4	3	2	2	−1	−3	−4	−3 456	0,7047	$-2{,}0390 \cdot 10^{-4}$
5	7	5	4	3	1	1	−2	−3	2 520	0,6703	$2{,}6530 \cdot 10^{-4}$
6	9	7	6	5	3	2	−1	−1	11 340	0,6065	$0{,}5348 \cdot 10^{-4}$
7	10	8	7	6	4	3	1	−2	−80 640	0,5769	$-0{,}0715 \cdot 10^{-4}$
$\Pi(9) = 3840$											$S = 1{,}6535 \cdot 10^{-4}$

On en tire

$$y(0{,}45) = \Pi(9) \sum_{i=0}^{i=7} \frac{y_i}{D_i} = \Pi(9) \cdot S = 3840 \cdot 1{,}6535 \cdot 10^{-4} = \underline{0{,}6349}.$$

Exemple 2. La fonction $y = \cos x$ est donnée par le tableau [5]

x	5,0	5,1	5,2	5,3
y	0,283662185	0,377977743	0,468516671	0,554374336
t	0	1	2	3
x	5,4	5,5	5,6	5,7
y	0,634692876	0,708669774	0,775565879	0,834712785
t	4	5	6	7

Trouver $\cos 5{,}347$.

S o l u t i o n. Effectuons le changement de variable suivant la formule

$$x = 0{,}1t + 5.$$

Les valeurs de la variable t relatives aux points d'interpolation seront alors 0, 1, 2, 3, 4, 5, 6, 7 et la valeur recherchée $x = 5{,}347$ deviendra $x = t = 3{,}47$. En tenant compte du fait que les points $t_i = i$ $(i = 0, 1, \ldots, 7)$ sont équidistants, les calculs peuvent se faire d'après le schéma donné ci-dessus (tableau 48).

Tableau 48

Schéma de calcul des coefficients de Lagrange pour le cas des points équidistants

i	x_i	y_i	$t-i$	$(-1)^{7-i}C_7^i$	$(-1)^{7-i}C_7^i \frac{y_i}{t-i}$
0	5,0	0,283662185	3,47	−1	−0,08174702
1	5,1	0,377977743	2,47	7	1,07119198
2	5,2	0,468516671	1,47	−21	−6,69309530
3	5,3	0,554374336	0,47	35	41,28319523
4	5,4	0,634692876	−0,53	−35	41,91368048
5	5,5	0,708669774	−1,53	21	−9,72684003
6	5,6	0,775565879	−2,53	−7	2,14583444
7	5,7	0,834712785	−3,53	1	−0,23646254
$\Pi = 42{,}8848749$				$S = 69{,}67575724$	

Le tableau 48 donne:

$$\Pi(3{,}47) = \sum_{i=0}^{7} (3{,}47 - i) = 42{,}8848749$$

et

$$S = \sum_{i=0}^{7} (-1)^{7-i} C_7^i \frac{y_i}{3{,}47 - i} = 69{,}67575724.$$

En vertu de la formule (7)

$$\cos 5{,}347 = \frac{1}{7!} \cdot \Pi(3{,}47) \cdot S = 0{,}592864312.$$

§ 14. Evaluation de l'erreur de la formule de Lagrange

Au § 12 nous avons construit pour la fonction $y = f(x)$ le polynôme de Lagrange $L_n(x)$ qui prend aux points $x_0, x_1, \ldots, x_n$ les valeurs données

$$y_0 = f(x_0),\ y_1 = f(x_1),\ \ldots,\ y_n = f(x_n).$$

Une question se pose de savoir : quelle est en d'autres points l'approximation du polynôme construit par rapport à la fonction $f(x)$, c'est-à-dire quelle est la grandeur du reste

$$R_n(x) = f(x) - L_n(x).$$

Pour déterminer cette approximation imposons à la fonction $y = f(x)$ des restrictions supplémentaires. Supposons que dans le domaine considéré $a \leqslant x \leqslant b$ de variation de x, qui contient les points d'interpolation, la fonction possède toutes les dérivées $f'(x)$, $f''(x)$, ..., $f^{(n+1)}(x)$ jusqu'à l'ordre $(n+1)$ y compris.

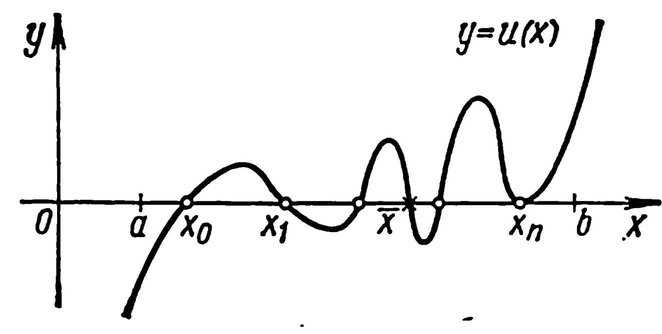


Fig. 63.

Introduisons une fonction auxiliaire

$$u(x) = f(x) - L_n(x) - k\Pi_{n+1}(x), \qquad (1)$$

où

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$$

et k est une constante qui sera choisie dans ce qui suit.

Il est évident que la fonction $u(x)$ possède $n+1$ racines aux points

$$x_0, x_1, \ldots, x_n.$$

Choisissons maintenant la constante k de sorte que $u(x)$ ait une $(n+2)$-ième racine en un point quelconque fixé $\bar{x}$ du segment $[a, b]$, autre que les points d'interpolation (fig. 63). A cet effet il suffit de poser

$$f(\bar{x}) - L_n(\bar{x}) - k\Pi_{n+1}(\bar{x}) = 0.$$

D'où, puisque $\Pi_{n+1}(\bar{x}) \neq 0$,

$$k = \frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})}. \qquad (2)$$

Pour cette valeur du coefficient k, la fonction $u(x)$ a $n+2$ racines sur le segment $[a, b]$ et s'annule aux extrémités de chaque segment

$$[x_0, x_1], [x_1, x_2], \ldots, [x_i, \bar{x}], [\bar{x}, x_{i+1}], \ldots, [x_{n-1}, x_n].$$

En appliquant le théorème de Rolle à chacun de ces segments on voit que la dérivée $u'(x)$ compte au moins $n+1$ racines sur le segment $[a, b]$. Opérant de même pour la dérivée $u'(x)$ on voit que la dérivée seconde $u''(x)$ devient nulle au moins n fois sur le segment $[a, b]$.

Finalement ces raisonnements aboutissent à la conclusion que sur le segment donné $[a, b]$, la dérivée $u^{(n+1)}(x)$ possède au moins un zéro que nous désignerons par ξ: $u^{(n+1)}(\xi) = 0$.

Comme

$$L_n^{(n+1)}(x)=0 \quad \text{et} \quad \Pi_{n+1}^{(n+1)}(x)=(n+1)\,!,$$

la formule (1) entraîne

$$u^{(n+1)}(x)=f^{(n+1)}(x)-k(n+1)\,!.$$

Pour $x=\xi$ on obtient:

$$0=f^{(n+1)}(\xi)-k(n+1)\,!.$$

D'où

$$k=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}\,. \tag{3}$$

La comparaison des seconds membres des formules (2) et (3) donne:

$$\frac{f(\bar{x})-L_n(\bar{x})}{\Pi_{n+1}(\bar{x})}=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}\,,$$

c'est-à-dire

$$f(\bar{x})-L_n(\bar{x})=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}\,\Pi_{n+1}(\bar{x}). \tag{4}$$

Puisque $\bar{x}$ est arbitraire, la formule (4) peut également s'écrire:

$$R_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)\,!}\,\Pi_{n+1}(x), \tag{5}$$

où ξ dépend de x et repose à l'intérieur du segment $[a, b]$.

Notons que la formule (5) est vraie pour tout point du segment $[a, b]$, y compris pour tout point d'interpolation.

Désignant par

$$M_{n+1}=\max_{a\leqslant x\leqslant b}|f^{(n+1)}(x)|,$$

on obtient l'estimation suivante de l'erreur absolue de la formule de Lagrange:

$$|R_n(x)|=|f(x)-L_n(x)|\leqslant\frac{M_{n+1}}{(n+1)\,!}\,|\Pi_{n+1}(x)|, \tag{6}$$

avec

$$\Pi_{n+1}(x) = (x-x_0)(x-x_1)\ldots(x-x_n). \tag{6'}$$

E x e m p l e. Avec quelle précision peut-on calculer $\sqrt{115}$ à l'aide de la formule de Lagrange pour la fonction $y = \sqrt{x}$ si l'on prend les points d'interpolation $x_0 = 100$, $x_1 = 121$, $x_2 = 144$?

S o l u t i o n. On a:

$$y' = \frac{1}{2} x^{-\frac{1}{2}}, \quad y'' = -\frac{1}{4} x^{-\frac{3}{2}}, \quad y''' = \frac{3}{8} x^{-\frac{5}{2}}.$$

Il en résulte

$$M_3 = \max |y'''| = \frac{3}{8} \cdot \frac{1}{\sqrt{100^5}} = \frac{3}{8} \cdot 10^{-5} \quad \text{pour} \quad 100 \leqslant x \leqslant 144.$$

En vertu de la formule (6)

$$|R_2| \leqslant \frac{3}{8} \cdot 10^{-5} \cdot \frac{1}{3!} |(115-100)\,(115-121)(115-144)| =$$

$$= \frac{1}{16} \cdot 10^{-5} \cdot 15 \cdot 6 \cdot 29 \approx 1{,}6 \cdot 10^{-3}.$$

§ 15. Evaluation des erreurs des formules de Newton

Si les points d'interpolation $x_0, x_1, \ldots, x_n$ sont équidistants et si

$$x_{i+1} - x_i = h \quad (i = 0, 1, 2, \ldots, n-1),$$

en posant

$$q = \frac{x - x_0}{h},$$

on obtient en vertu de la formule (5) du paragraphe précédent le *reste de la première formule de Newton*

$$R_n(x) = h^{n+1} \cdot \frac{q(q-1)\ldots(q-n)}{(n+1)!} f^{(n+1)}(\xi), \tag{1}$$

où ξ est une certaine valeur intermédiaire entre les points d'interpolation $x_0, x_1, \ldots, x_n$ et le point concerné x. Notons que dans le cas de l'interpolation au sens strict, $\xi \in [x_0, x_n]$; dans le cas de l'extrapolation, il est possible que $\xi \bar{\in} [x_0, x_n]$.

D'une façon analogue, en posant dans la formule (5) du § 14

$$q = \frac{x - x_n}{h},$$

on obtient le *reste de la deuxième formule de Newton*

$$R_n(x) = h^{n+1} \cdot \frac{q(q+1)\ldots(q+n)}{(n+1)!} f^{(n+1)}(\xi), \tag{2}$$

où ξ est une valeur intermédiaire entre les points d interpolation $x_0, x_1, \ldots, x_n$ et le point x.

Dans les cas pratiques il est d'usage d'arrêter le calcul suivant les formules de Newton aux termes qui contiennent des différences

pouvant être considérées comme constantes dans les limites de la précision imposée.

En supposant que $\Delta^{n+1} y$ soient quasi constantes pour la fonction $y = f(x)$ et h suffisamment petit, et en tenant compte du fait que

$$f^{(n+1)}(x) = \lim_{h \to 0} \frac{\Delta^{n+1} y}{h^{n+1}},$$

on peut poser approximativement :

$$f^{(n+1)}(\xi) \approx \frac{\Delta^{n+1} y_0}{h^{n+1}}.$$

Dans ce cas, le reste de la première formule de Newton est égal à

$$R_n(x) \approx \frac{q(q-1)\ldots(q-n)}{(n+1)!} \Delta^{n+1} y_0.$$

Sous ces mêmes conditions, pour le reste de la deuxième formule de Newton on obtient l'expression

$$R_n(x) \approx \frac{q(q+1)\ldots(q+n)}{(n+1)!} \Delta^{n+1} y_n.$$

Exemple 1. Les tables des logarithmes à cinq décimales donnent les logarithmes des entiers de $x = 1\,000$ à $x = 10\,000$ avec une borne d'erreur absolue égale à $\frac{1}{2} \cdot 10^{-5}$. Est-il possible de réaliser une interpolation linéaire avec la même précision?

Solution. Posant

$$y = \lg x$$

on aura :

$$y' = \frac{M}{x} \quad \text{et} \quad y'' = -\frac{M}{x^2},$$

avec $M = 0{,}43$. D'où

$$M_2 = \max |y''| < \frac{0{,}5}{10^6} = \frac{1}{2} \cdot 10^{-6}.$$

Pour $n = 2$ et $h = 1$ la formule (1) donne l'estimation suivante de l'erreur d'interpolation linéaire :

$$|R_1(x)| \leqslant \frac{|q(q-1)|}{2!} M_2 \leqslant \frac{q(1-q)}{2} \cdot \frac{1}{2} \cdot 10^{-6}.$$

Comme pour $0 \leqslant q \leqslant 1$ on a

$$q(1-q) = \frac{1}{4} - \left(\frac{1}{2} - q\right)^2 \leqslant \frac{1}{4},$$

on obtient finalement :

$$|R_1(x)| \leqslant \frac{\frac{1}{4}}{2} \cdot \frac{1}{2} \cdot 10^{-6} < 10^{-7}.$$

Par conséquent, l'interpolation linéaire est tout à fait admissible.

Exemple 2. Evaluer l'erreur d'approximation de la fonction $f(x) = \sin x$ par le polynôme de cinquième degré $P_5(x)$ coïncidant avec la fonction donnée pour les valeurs $x = 0°, 5°, 10°, 15°, 20°, 25°$.

Solution. Ici $f^{(6)}(x) = -\sin x$; par suite $|f^{(6)}(x)| \leqslant 1$. En vertu de la formule (1), on a:

$$|\sin x - P_5(x)| \leqslant$$
$$\leqslant \frac{1}{6!}\left| x\left(x-\frac{\pi}{36}\right)\left(x-\frac{\pi}{18}\right)\left(x-\frac{\pi}{12}\right)\left(x-\frac{\pi}{9}\right)\left(x-\frac{5\pi}{36}\right)\right|.$$

Par exemple, pour $x = 12°30' = \text{arc } 0{,}21816$, on obtient:

$$|\sin x - P_5(x)| < 2{,}2 \cdot 10^{-9}.$$

§ 16. Evaluation des erreurs des formules d'interpolation par différences centrales

Voici sans démonstration les expressions du reste des formules de Stirling et de Bessel [3].

a) *Reste de la formule d'interpolation de Stirling.* Si $2n$ est l'ordre maximal des différences utilisées du tableau et si $x \in [x_0 - nh, x_0 + nh]$, alors

$$R_n(x) = \frac{h^{2n+1} f^{(2n+1)}(\xi)}{(2n+1)!} q(q^2-1^2)(q^2-2^2)(q^2-3^2)\ldots(q^2-n^2),$$

où

$$q = \frac{x - x_0}{h} \quad \text{et} \quad \xi \in [x_0 - nh, x_0 + nh].$$

Si l'expression analytique de la fonction $f(x)$ est inconnue, on pose pour un h petit:

$$R_n(x) \approx \frac{\Delta^{2n+1}y_{-n-1} + \Delta^{2n+1}y_{-n}}{2(2n+1)!} q(q^2-1^2)(q^2-2^2)\ldots(q^2-n^2).$$

b) *Reste de la formule d'interpolation de Bessel.* Si $2n+1$ est l'ordre de la différence maximale utilisée du tableau et si $x \in [x_0 - nh, x_0 + (n+1)h]$, alors

$$R_n(x) = \frac{h^{2n+2}}{(2n+2)!} f^{(2n+2)}(\xi)\, q(q^2-1^2)(q^2-2^2) \times \ldots$$
$$\ldots \times (q^2-n^2)[q-(n+1)],$$

où

$$q = \frac{x - x_0}{h} \quad \text{et} \quad \xi \in [x_0 - nh, x_0 + (n+1)h].$$

Si encore la fonction $f(x)$ est donnée par le tableau et le pas h est petit, on adopte:

$$R_n(x) \approx \frac{\Delta^{2n+2}y_{-n-1} + \Delta^{2n+2}y_{-n}}{2(2n+2)!} q(q^2-1^2)(q^2-2^2) \times \ldots$$
$$\ldots \times (q^2-n^2)[q-(n+1)].$$

En particulier, avec $q=\frac{1}{2}$ une *erreur de la formule de dichotomie* s'écrit

$$R_n=\frac{h^{2n+2}f^{(2n+2)}(\xi)}{(2n+2)!}(-1)^{n+1}\frac{[1\cdot 3\cdot 5\ldots(2n+1)]^2}{2^{2n+2}}$$

ou

$$R_n\approx\frac{\Delta^{2n+2}y_{-n-1}+\Delta^{2n+2}y_{-n}}{2(2n+2)!}(-1)^{n+1}\frac{[1\cdot 3\cdot 5\ldots(2n+1)]^2}{2^{2n+2}}.$$

Si l'on pose

$$q=p+\frac{1}{2},$$

l'expression du reste de la formule de Bessel se met sous la forme

$$R_n(x)=\frac{h^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi)\left(p^2-\frac{1}{4}\right)\left(p^2-\frac{9}{4}\right)\ldots\left[p^2-\frac{(2n+1)^2}{4}\right].$$

§ 17. Sur le meilleur choix des points d'interpolation

L'analyse de la formule (5) du § 14 montre que l'erreur $R_n(x)$ de la formule de Lagrange est, à une constante numérique près, le produit de deux facteurs, dont l'un, $f^{(n+1)}(\xi)$, dépend des propriétés de la fonction $f(x)$ et ne se prête pas à l'ajustage, alors que la grandeur de l'autre, $\Pi_{n+1}(x)$, n'est déterminée que par le choix des points d'interpolation.

Dans le cas d'une mauvaise répartition des points d'interpolation x_i, la borne supérieure du module de l'erreur $R_n(x)$ ((6) du § 14) peut être très grande. Par exemple, si les points x_i se concentrent au voisinage de l'une des extrémités du segment $[a, b]$, $R_n(x)$ sera dans le cas général grand aux points x proches de l'autre extrémité du segment. Le choix des points d'interpolation x_i (pour le nombre n donné de points) doit être donc le meilleur pour que le polynôme $\Pi_{n+1}(x)$ soit sur le segment $[a, b]$ minimal en valeur absolue maximale, ou comme on dit pour abréger, « s'écarte de zéro sur $[a, b]$ le moins possible ». Ce problème a été résolu par le mathématicien russe P. Tchébychev [2], [6] qui a montré que dans ce sens le meilleur choix des points d'interpolation est donné par la formule

$$x_i=\frac{b+a}{2}+\frac{b-a}{2}\xi_i,$$

où

$$\xi_i=-\cos\frac{2i+1}{2n+2}\pi\quad(i=0,\ 1,\ 2,\ \ldots,\ n)$$

sont les zéros du polynôme dit de Tchébychev $T_{n+1}(x)$. Dans ce cas on a:

$$|\Pi_{n+1}(x)|\leqslant 2\left(\frac{b-a}{4}\right)^{n+1}.$$

Il est curieux de noter que ces points ne sont pas équidistants, mais s'accumulent aux extrémités du segment. Même avec un tel choix de points on ne peut pas garantir dans le cas général que la valeur absolue de l'erreur soit aussi petite que l'on veut pour un n suffisamment grand.

Voici des remarques générales sur la détermination des erreurs des formules d'interpolation. Si les différences maximales sont pratiquement constantes, le résultat d'interpolation au sens strict compte généralement autant de décimales exactes qu'en comptent les données tabulées; l'évaluation des erreurs n'est donc pas obligatoire. Dans le cas de la formule de Lagrange, il est impossible de suivre la variation des différences et, par suite, il faut, si possible, évaluer le reste.

Si la fonction $f(x)$ est tabulée et que son expression analytique soit inconnue, l'évaluation de l'erreur du polynôme d'interpolation est en toute rigueur impossible. En effet, théoriquement on peut construire pour le polynôme considéré un nombre infini de fonctions différentes coïncidant avec ce polynôme dans le système de points donné. Ainsi aux points intermédiaires, l'écart du polynôme d'interpolation par rapport à la fonction peut être quelconque. Toutefois, si la fonction est telle que sa courbe est régulière, les erreurs des polynômes d'interpolation peuvent être déterminées approximativement avec un grand degré de certitude à partir des valeurs des différences d'ordres supérieurs d'après les formules données dans ce qui précède.

§ 18. Différences divisées

Jusqu'à présent, en dressant le tableau des différences, nous avons supposé que les valeurs de l'argument d'une fonction sont équidistantes, c'est-à-dire que leur *pas* est *constant*. Toutefois dans la pratique on rencontre également des tableaux pour des valeurs non équidistantes de l'argument, c'est-à-dire des tableaux au pas variable. Il en est souvent ainsi, par exemple, des données empiriques. Pour les tableaux au pas variable, la notion des différences finies est généralisée, et on introduit ce qu'on appelle les *différences divisées*.

Supposons que la fonction $y = f(x)$ soit tabulée, $x_0, x_1, x_2, \ldots$ les valeurs de son argument et $y_0, y_1, y_2, \ldots$ les valeurs respectives de la fonction où les différences

$$\Delta x_i = x_{i+1} - x_i \neq 0 \qquad (i = 0, 1, \ldots)$$

ne sont pas égales entre elles.

Les relations

$$[x_i, x_{i+1}] = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}$$

($i = 0, 1, 2, \ldots$) s'appellent *différences premières divisées.* Par exemple,

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0}; \quad [x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1}, \quad \text{etc.}$$

D'une façon analogue on détermine les *différences secondes divisées*

$$[x_i, x_{i+1}, x_{i+2}] = \frac{[x_{i+1}, x_{i+2}] - [x_i, x_{i+1}]}{x_{i+2} - x_i}$$

($i = 0, 1, 2, \ldots$). Par exemple,

$$[x_0, x_1, x_2] = \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0},$$

etc.

D'une façon générale, les *différences divisées d'ordre n* s'obtiennent à partir des différences divisées d'ordre ($n - 1$) à l'aide de la relation récurrente

$$[x_i, x_{i+1}, \ldots, x_{i+n}] = \frac{[x_{i+1}, \ldots, x_{i+n}] - [x_i, \ldots, x_{i+n-1}]}{x_{i+n} - x_i} \qquad (1)$$

($n = 1, 2, \ldots$; $i = 0, 1, 2, \ldots$).

Remarquons que les différences divisées ne changent pas avec la permutation des éléments, c'est-à-dire qu'elles sont des *fonctions symétriques* de leurs arguments. Par exemple,

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{y_0 - y_1}{x_0 - x_1} = [x_1, x_0], \quad \text{etc.}$$

Les différences divisées forment généralement un tableau du type suivant (tableau 49).

Tableau 49

Différences divisées

x	y	Différences divisées			
		ordre 1	ordre 2	ordre 3	ordre 4
x_0	y_0				
		$[x_0, x_1]$			
x_1	y_1		$[x_0, x_1, x_2]$		
		$[x_1, x_2]$		$[x_0, x_1, x_2, x_3]$	
x_2	y_2		$[x_1, x_2, x_3]$		$[x_0, x_1, x_2, x_3, x_4]$
		$[x_2, x_3]$		$[x_1, x_2, x_3, x_4]$	
x_3	y_3		$[x_2, x_3, x_4]$		
		$[x_3, x_4]$			
x_4	y_4				

Exemple. Composer les différences divisées de la fonction donnée par le tableau suivant:

x	0	0,2	0,3	0,4	0,7	0,9
y	132,651	148,877	157,464	166,375	195,112	216,000

Solution. En appliquant successivement la formule (1), on aura:

$$[x_0, x_1] = \frac{148,877-132,651}{0,2-0} = 81,13;$$

$$[x_1, x_2] = \frac{157,464-148,877}{0,3-0,2} = 85,87;$$

$$[x_0, x_1, x_2] = \frac{85,87-81,13}{0,3-0} = 15,8,$$

etc. Les résultats des calculs sont portés sur le tableau 50.

Tableau 50

Différences divisées de la fonction y

x	y	ordre 1	ordre 2	ordre 3	ordre 4
0	132,651				
		81,13			
0,2	140,877		15,8		
		85,87		1	
0,3	157,464		16,2		0
		89,11		1	
0,4	166,375		16,7		0
		95,79		1	
0,7	195,112		17,3		
		104,44			
0,9	216,000				

§ 19. Formule de Newton pour des valeurs non équidistantes de l'argument

En utilisant la notion des différences divisées on peut mettre la formule de Lagrange sous une forme analogue à la première formule de Newton. Démontrons au préalable un lemme qui présente lui-même un intérêt propre.

L e m m e. *Si $y = P(x)$ est un polynôme de degré n, sa différence divisée d'ordre $(n+1)$ est identiquement nulle, c'est-à-dire*

$$[x, x_0, x_1, \ldots, x_n] \equiv 0$$

pour un système quelconque des nombres distincts entre eux x, x_0, $x_1, \ldots, x_n$.

En effet, si $P(x)$ est un polynôme de degré n,

$$[x, x_0] = \frac{P(x) - P(x_0)}{x - x_0} \equiv P(x, x_0)$$

est un polynôme en x de degré $(n-1)$. Ensuite

$$[x, x_0, x_1] = \frac{P(x, x_0) - P(x_0, x_1)}{x - x_1} \equiv P(x, x_0, x_1)$$

est un polynôme en x de degré $(n-2)$. En effet, la fonction $P(x, x_0) - P(x_0, x_1) = P(x, x_0) - P(x_1, x_0)$ admet pour racine $x = x_1$ et, par conséquent, en vertu du théorème de Bézout, le polynôme $P(x, x_0) - P(x_0, x_1)$ est divisible par le binôme $x - x_1$. Des raisonnements analogues montrent que

$$[x, x_0, \ldots, x_{n-1}] \equiv P(x, x_0, \ldots, x_{n-1})$$

est un polynôme de degré nul, c'est-à-dire

$$P(x, x_0, \ldots, x_{n-1}) = C.$$

D'où

$$[x, x_0, \ldots, x_n] = \frac{C - C}{x - x_n} \equiv 0.$$

Soit maintenant $P(x)$ un polynôme de Lagrange de degré n tel que

$$P(x_i) = f(x_i) = y_i \tag{1}$$

$(i = 0, 1, \ldots, n)$, où $y = f(x)$ est la fonction donnée. Désignons par $P(x, x_0)$, $P(x, x_0, x_1)$, $\ldots$, $P(x, x_0, \ldots, x_n)$ les différences divisées successives du polynôme $P(x)$. On a :

$$\left.\begin{aligned} &P(x_0, x_1) = [x_0, x_1], \\ &P(x_0, x_1, x_2) = [x_0, x_1, x_2], \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &P(x_0, x_1, \ldots, x_n) = [x_0, x_1, \ldots, x_n]\,; \end{aligned}\right\} \tag{2}$$

de plus, en vertu du lemme,

$$P(x, x_0, \ldots, x_n) = 0. \tag{3}$$

Par définition

$$\frac{P(x) - P(x_0)}{x - x_0} = P(x, x_0)\,; \tag{4}$$

d'où

$$P(x) = P(x_0) + P(x, x_0)(x - x_0). \tag{5}$$

Par définition

$$P(x, x_0, \ldots, x_m) = \frac{P(x, x_0, \ldots, x_{m-1}) - P(x_0, \ldots, x_m)}{x - x_m}.$$

On en tire

$$P(x, x_0, \ldots, x_{m-1}) = P(x_0, \ldots, x_m) + \\ + (x - x_m) P(x, x_0, \ldots, x_m) \tag{6}$$

$(m = 1, 2, \ldots, n)$.

Utilisant la formule (6) on déduit, de proche en proche, de la formule (5):

$$\begin{aligned} P(x) = P(x_0) + P(x, x_0)(x - x_0) = \\ &= P(x_0) + P(x_0, x_1)(x - x_0) + \\ &+ P(x, x_0, x_1)(x - x_0)(x - x_1) = \\ &= P(x_0) + P(x_0, x_1)(x - x_0) + \\ &+ P(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots \\ &\ldots + P(x_0, x_1, \ldots, x_n)(x - x_0)(x - x_1) \ldots \\ &\ldots (x - x_{n-1}) + P(x, x_0, \ldots, x_n) \times \\ &\times (x - x_0)(x - x_1) \ldots (x - x_n), \end{aligned}$$

ou, en tenant compte des égalités (2) et (3), finalement on obtient la *formule de Newton pour les valeurs d'argument non équidistantes*

$$P(x) = y_0 + [x_0, x_1](x - x_0) + [x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots + \\ + [x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \ldots (x - x_{n-1}). \tag{7}$$

Comme dans les cas courants, l'erreur de la formule (7) s'écrit

$$R(x) = f(x) - P(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1) \ldots (x - x_n), \tag{8}$$

où ξ est une valeur intermédiaire entre les points $x_0, x_1, \ldots, x_n$ et x.

Exemple. Former le polynôme d'interpolation de la fonction $y = f(x)$ donnée par le tableau:

x	0	2,5069	5,0154	7,52270
y	0,3989423	0,3988169	0,3984408	0,3978138

Trouver à l'aide de ce polynôme f (3,7608).

S o l u t i o n. Calculons les différences divisées de la fonction y (tableau 51).

Tableau 51

Différences divisées de la fonction y

x	y	ordre 1	ordre 2	ordre 3
0	0,3989423			
		—500		
2,5069	0,3988169		—199	
		—1499		0
5,0154	0,3984408		—199	
		—2496		
7,5270	0,3978138			

En utilisant la formule (7), on tombe sur

$$y = 0{,}3989423 - 0{,}0000500x - 0{,}0000199x\,(x - 2{,}5069).$$

D'où

$$y\,(3{,}7608) = 0{,}3989423 - 0{,}0000500 \cdot 3{,}7608 - 0{,}0000199 \cdot 3{,}7608 \cdot (3{,}7608 - 2{,}5069) = 0{,}3986604.$$

§ 20. Interpolation inverse pour le cas des points équidistants

Soit la fonction $y = f\,(x)$ donnée par le tableau.

La tâche de l'*interpolation inverse* consiste à calculer d'après la valeur donnée de la fonction y la valeur correspondante de l'argument x.

Considérons d'abord le cas des points équidistants. A cette fin on recourt d'ordinaire à la *méthode des approximations successives*.

Supposons que la fonction $y = f\,(x)$ soit monotone et que la valeur donnée de y est comprise entre $y_0 = f\,(x_0)$ et $y_1 = f\,(x_1)$.

En remplaçant la fonction y par le premier polynôme de Newton, on obtient :

$$y = y_0 + \frac{\Delta y_0}{1\,!}\,q + \frac{\Delta^2 y_0}{2\,!}\,q\,(q-1) + \ldots + \frac{\Delta^n y_0}{n\,!}\,q\,(q-1)\ldots(q-n+1)\,;$$

d'où $q = \varphi\,(q)$, avec

$$\varphi\,(q) = \frac{y-y_0}{\Delta y_0} - \frac{\Delta^2 y_0}{2\,!\Delta y_0}\,q\,(q-1) - \ldots - \frac{\Delta^n y_0}{n\,!\Delta y_0}\,q\,(q-1)\ldots(q-n+1).$$

On admet pour approximation initiale:

$$q_0 = \frac{y - y_0}{\Delta y_0}.$$

En appliquant la méthode des approximations successives, on aura:

$$q_m = \varphi(q_{m-1}) \qquad (m = 1, 2, \ldots). \qquad (1)$$

Si $f(x) \in C^{(n+1)}[a, b]$, où l'intervalle $[a, b]$ contient les points d'interpolation et que le pas h soit suffisamment petit, ce processus converge, c'est-à-dire

$$\lim_{m \to \infty} q_m = q,$$

où q est une solution réelle.

Pratiquement le processus itératif se poursuit tant que les chiffres de précision imposée ne deviennent invariables et on pose $q \approx \approx q_s$, où q_s est la dernière approximation.

Après avoir trouvé q on détermine x suivant la formule

$$\frac{x - x_0}{h} = q;$$

d'où

$$x = x_0 + qh.$$

Exemple 1. Utiliser les valeurs de la fonction $y = \lg x$ données par le tableau

x	20	25	30
y	1,3010	1,3979	1,4771

pour trouver la valeur de x telle que $y = \lg x = 1,35$.

Solution. Formons le tableau des différences.

Tableau 52

Différences de la fonction y

x	y	Δy	$\Delta^2 y$
20	1,3010	969	−177
25	1,3979	792	
30	1,4771		

Adoptant $y_0 = 1{,}3010$, on aura :

$$q_0 = \frac{y - y_0}{\Delta y_0} = \frac{1{,}35 - 1{,}3010}{0{,}0969} = \frac{490}{969} = 0{,}506.$$

Ensuite, en gardant trois décimales, on a par le procédé de proche en proche :

$$q_1 = 0{,}506 - \frac{177}{2 \cdot 969} \cdot 0{,}506\,(1 - 0{,}506) = 0{,}506 - 0{,}023 = 0{,}483\,;$$

$$q_2 = 0{,}506 - \frac{177}{2 \cdot 969} \cdot 0{,}483\,(1 - 0{,}483) = 0{,}506 - 0{,}023 = 0{,}483.$$

On pose

$$q = 0{,}483.$$

D'où

$$x = x_0 + qh = 20 + 0{,}483 \cdot 5 = 22{,}42.$$

Le tableau des antilogarithmes donne $x = 22{,}39$. L'écart important entre les valeurs calculée et exacte s'explique par le fait que le pas $h = 5$ est trop grand.

Nous avons appliqué la méthode des approximations successives à la résolution d'un problème d'interpolation inverse en recourant à la première formule de Newton. Mais d'une façon tout à fait analogue on peut l'appliquer également à d'autres formules d'interpolation, et notamment à la deuxième formule de Newton, aux formules de Stirling, de Bessel, etc. Illustrons ce fait par l'exemple suivant.

E x e m p l e 2. Le tableau 53 donne les valeurs de l'intégrale de probabilité [3]

$$y = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2}\, dx.$$

Pour quelle valeur de x l'intégrale y est-elle égale à $\frac{1}{2}$?

Tableau 53

Valeurs de l'intégrale de probabilité

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0,45	0,4754818				
		91737			
0,46	0,4846555		−840		
		90897		−11	
0,47	0,4937452		−851		1
		90046		−10	
0,48	0,5027498		−861		2
		89185		−8	
0,49	0,5116683		−869		
		88316			
0,50	0,5204999				

S o l u t i o n. Complétons le tableau 53 par les différences de la fonction y. La grandeur tabulée la plus proche de l'argument x, associée à la valeur de la fonction $y = \frac{1}{2}$, est $x_0 = 0{,}47$. Ici il est commode d'employer la formule de Bessel.

On a $x_0 = 0{,}47$; $h = 0{,}01$; $y = 0{,}5$.

En portant ces valeurs dans la formule (8) du § 7 et en utilisant les données tabulées correspondantes, on obtient:

$$0{,}5 = 0{,}4982475 + 0{,}0090046p + \frac{p^2 - 0{,}25}{2}\left(\frac{-851-861}{2}\right)\cdot 10^{-7} + \\ + \frac{p(p^2-0{,}25)}{6}(-10)\cdot 10^{-7}. \quad (2)$$

On en tire en divisant les deux membres de l'égalité (2) par 0,0090046 et en isolant le terme de premier degré en p:

$$p = 0{,}194623 + 4{,}753\cdot 10^{-3}(p^2 - 0{,}25) + \\ + 1{,}85\cdot 10^{-5}p(p^2 - 0{,}25). \quad (3)$$

Admettons pour première approximation du paramètre p:

$$p^{(1)} = 0{,}194623.$$

Portant $p^{(1)}$ dans l'expression (3) on obtient la deuxième approximation:

$$p^{(2)} = 0{,}194623 + 4{,}753\cdot 10^{-3}[(0{,}194623)^2 - 0{,}25] + \\ + 1{,}85\cdot 10^{-5}\cdot 0{,}194623\cdot[(0{,}194623)^2 - 0{,}25] = \\ = 0{,}194623 - 0{,}001008 - 0{,}000001 = 0{,}193614.$$

De même, en portant dans la formule (3) $p^{(2)}$ au lieu de p, on obtient la troisième approximation:

$$p^{(3)} = 0{,}193612.$$

Comme les premières cinq décimales coïncident, le processus itératif peut être considéré comme achevé.

Ensuite on trouve successivement:

$$q = p + \frac{1}{2} = 0{,}693612$$

et

$$x = x_0 + qh = 0{,}47 + 0{,}01\cdot 0{,}693612 = 0{,}47693612.$$

Les six premières décimales de cette valeur sont exactes.

§ 21. Interpolation inverse pour le cas des points non équidistants

Le problème d'interpolation inverse pour le cas des valeurs non équidistantes de l'argument $x_0, x_1, \ldots, x_n$ peut être résolu immédiatement à l'aide de la formule de Lagrange. A cette fin il suffit d'admettre que la variable y est indépendante et écrire la formule qui exprime x en fonction de y (fig. 64)

$$x=\sum_{i=0}^{n}\frac{(y-y_1)(y-y_2)\ldots(y-y_{i-1})(y-y_{i+1})\ldots(y-y_n)}{(y_i-y_1)(y_i-y_2)\ldots(y_i-y_{i-1})(y_i-y_{i+1})\ldots(y_i-y_n)}x_i, \quad (1)$$

où $y_i = f(x_i)$ $(i = 0, 1, \ldots, n)$. On peut également, en considérant y comme argument, utiliser la formule de Newton pour les valeurs non équidistantes de l'argument (cf. § 19):

$$x = x_0 + [y_0, y_1](y - y_0) + [y_0, y_1, y_2](y - y_0)(y - y_1) + \ldots + [y_0, y_1, \ldots, y_n] \times (y - y_0)(y - y_1)\ldots(y - y_{n-1}), \quad (2)$$

où $[y_0, y_1]$, $[y_0, y_1, y_2]$, ..., $[y_0, y_1, \ldots, y_n]$ sont les différences divisées correspondantes.

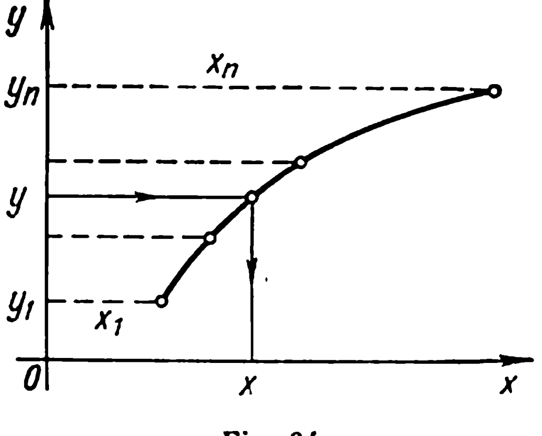


Fig. 64.

Exemple. Résoudre l'exemple 2 du § 20 à l'aide de la formule de Lagrange pour l'interpolation inverse [3].

Solution. Bornons-nous aux quatre valeurs:

$$x_0 = 0{,}46; \quad x_1 = 0{,}47; \quad x_2 = 0{,}48; \quad x_4 = 0{,}49.$$

En posant

$$u = 10^7 y - \frac{1}{2}\cdot 10^7,$$

on aura le tableau suivant:

x	0,46	0,47	0,48	0,49
u	−153445	−62548	27498	116683

La valeur donnée $y = \frac{1}{2}$ correspond à $u = 0$. En appliquant la formule (2), où y est remplacé par u, on obtient:

$$x = \frac{62548\cdot(-27498)\cdot(-116683)}{(-153445+62548)\cdot(-153445-27498)\cdot(-153445-116683)}\cdot 0{,}46 +$$
$$+\frac{153445\cdot(-27498)(-116683)}{(-62548+153445)\cdot(-62548-27498)\cdot(-62548-116683)}\cdot 0{,}47 +$$
$$+\frac{153445\cdot 62548\cdot(-116683)}{(27498+153445)\cdot(27498+62548)\cdot(27498-116683)}\cdot 0{,}48 +$$
$$+\frac{153445\cdot 62548\cdot(-27498)}{(116683+153445)\cdot(116683+62548)\cdot(116683-27498)}\cdot 0{,}49 =$$
$$= -0{,}020779 + 0{,}157737 + 0{,}369928 - 0{,}029950 = 0{,}476936.$$

§ 22. Recherche des racines d'une équation par la méthode d'interpolation inverse

Notons en conclusion que la résolution de l'équation

$$f(x) = 0$$

peut être ramenée au problème d'interpolation inverse. A cette fin il faut dresser le tableau des valeurs de la fonction $y = f(x)$ et construire le tableau correspondant des différences pour les valeurs de x voisines de la racine. Ensuite on applique les procédés d'interpolation inverse en recherchant la valeur de x associée à $y = 0$.

Exemple. Trouver à 10^{-3} près à partir du tableau des valeurs de la fonction de Bessel $y = J_0(x)$ la racine de l'équation $J_0(x) = 0$ comprise dans l'intervalle (2,4; 2,6).

x	2,4	2,5	2,6
y	0,0025	−0,0484	−0,0968

Solution. Formons le tableau des différences (tableau 54). Adoptons $y = 0$ et $x_0 = 2{,}4$; $y_0 = 0{,}0025$; on obtient alors en vertu de la formule (1) du § 20:

$$q_0 = \frac{y - y_0}{\Delta y_0} = \frac{0{,}0025}{0{,}0509} = 0{,}049;$$

$$q_1 = q_0 + \frac{\Delta^2 y_0}{2\Delta y_0}\, q_0(1 - q_0) =$$
$$= 0{,}049 - \frac{25}{2\cdot 509}\cdot 0{,}049\cdot 0{,}951 = 0{,}049 - 0{,}001 = 0{,}048;$$

$$q_2 = 0{,}049 - \frac{25}{2\cdot 509}\cdot 0{,}048\cdot 0{,}952 = 0{,}049 - 0{,}001 = 0{,}048.$$

Tableau 54

Différences de la fonction de Bessel $y = J_0(x)$

x	y	Δy	$\Delta^2 y$
2,4	0,0025	−509	25
2,5	−0,0484	−484	
2,6	−0,0968		

Adoptons

$$q = 0{,}048;$$

d'où

$$x = x_0 + qh = 2{,}4 + 0{,}048 \cdot 0{,}1 = 2{,}405.$$

Les tables donnent

$$x = 2{,}4048.$$

§ 23. Méthode d'interpolation pour développer le déterminant caractéristique

L'interpolation des fonctions peut être utilisée pour développer le déterminant caractéristique (séculaire) (cf. chapitre XII)

$$D(\lambda) = \det(A - \lambda E),$$

où $A = [a_{ij}]$.

Choisissons des points équidistants

$$\lambda_0 = 0, \quad \lambda_1 = 1, \ldots, \quad \lambda_n = n$$

et calculons pour le déterminant $D(\lambda)$ les valeurs correspondantes

$$D(0) = D_0, \quad D(1) = D_1, \ldots, \quad D(n) = D_n.$$

En dressant le tableau horizontal des différences de la suite des nombres $D(0)$, $D(1)$, ..., $D(n)$, on trouve par le procédé usuel les différences $\Delta^i D(0)$ $(i = 0, 1, \ldots, n)$. D'où l'on tire, en appliquant la première formule de Newton, l'expression polynomiale du déterminant caractéristique

$$D(\lambda) = D(0) + \sum_{i=1}^{n} \frac{\Delta^i D(0)}{i!} \lambda(\lambda - 1) \ldots (\lambda - i + 1). \qquad (1)$$

Si l'on pose

$$\frac{\lambda(\lambda-1)\ldots(\lambda-i+1)}{i!}=\sum_{m=1}^{i} c_{mi}\lambda^m \quad (i=1, 2, \ldots), \tag{2}$$

après des transformations élémentaires on obtient la *formule de Markov*:

$$D(\lambda)=D(0)+\sum_{m=1}^{n}\lambda^m\sum_{i=m}^{n}c_{mi}\Delta^i D(0). \tag{3}$$

Les calculs suivant la formule (2) sont rendus plus faciles par les tableaux des coefficients c_{mi} [8].

Dans un cas plus général, si l'on prend comme points d'interpolation les nombres $\lambda_i = a + ih$ $(i = 0, 1, \ldots, n)$, la formule (3) s'écrit

$$D(\lambda)=D(a)+\sum_{m=1}^{n}(\lambda-a)^m\sum_{i=m}^{n}c_{mi}h^i\Delta^i D(a). \tag{4}$$

Bien que la *méthode d'interpolation* qui vient d'être exposée impose de longs calculs de $n+1$ déterminants d'ordre n, cette méthode est pour autant commode par son schéma de calcul très simple. De plus, elle est applicable au développement d'un déterminant de forme plus générale

$$F(\lambda) = \det[f_{ij}(\lambda)],$$

où $f_{ij}(\lambda)$ sont des polynômes entiers en λ.

Exemple. En utilisant la méthode d'interpolation développer le déterminant caractéristique

$$D(\lambda)=\begin{vmatrix} 1-\lambda & 2 & 3 & 4 \\ 2 & 1-\lambda & 2 & 3 \\ 3 & 2 & 1-\lambda & 2 \\ 4 & 3 & 2 & 1-\lambda \end{vmatrix}$$

(cf. chapitre XII, § 3, exemple).

Solution. Calculons successivement $D(i)$ pour $i = 0, 1, 2, 3, 4$. On a:

$$D(0) = -20,\quad D(1) = -119,\quad D(2) = -308,$$

$$D(3) = -575,\quad D(4) = -884.$$

Les différences $\Delta^i D(0)$ $(i = 0, 1, 2, 3, 4)$ sont consignées sur le tableau 55.

Tableau 55

Différences des nombres $D(\lambda)$

λ	$D(\lambda)$	$\Delta D(\lambda)$	$\Delta^2 D(\lambda)$	$\Delta^3 D(\lambda)$	$\Delta^4 D(\lambda)$
0	−20	−99	−90	12	24
1	−119	−189	−78	36	
2	−308	−267	−42		
3	−575	−309			
4	−884				

Puisque

$$\frac{\lambda}{1!} = \lambda;$$

$$\frac{\lambda(\lambda-1)}{2!} = \frac{\lambda^2}{2} - \frac{\lambda}{2};$$

$$\frac{\lambda(\lambda-1)(\lambda-2)}{3!} = \frac{\lambda^3}{6} - \frac{\lambda^2}{2} + \frac{\lambda}{3};$$

$$\frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)}{4!} = \frac{\lambda^4}{24} - \frac{\lambda^3}{4} + \frac{11\lambda^2}{24} - \frac{\lambda}{4},$$

la formule (2) conduit à

$$c_{11} = 1;$$

$$c_{22} = \frac{1}{2}, \quad c_{12} = -\frac{1}{2};$$

$$c_{33} = \frac{1}{6}, \quad c_{23} = -\frac{1}{2}, \quad c_{13} = \frac{1}{3};$$

$$c_{44} = \frac{1}{24}, \quad c_{34} = -\frac{1}{4}, \quad c_{24} = \frac{11}{24}, \quad c_{14} = -\frac{1}{4}.$$

D'où, en appliquant la formule de Markov (3),

$$D(\lambda) = D(0) + [c_{11}\Delta D(0) + c_{12}\Delta^2 D(0) + c_{13}\Delta^3 D(0) +$$
$$+ c_{14}\Delta^4 D(0)]\,\lambda + [c_{22}\Delta^2 D(0) + c_{23}\Delta^3 D(0) + c_{24}\Delta^4 D(0)]\,\lambda^2 +$$
$$+ [c_{33}\Delta^3 D(0) + c_{34}\Delta^4 D(0)]\,\lambda^2 + c_{44}\Delta^4 D(0)\,\lambda^4 =$$
$$= -20 + \left(-99\cdot 1 + 90\cdot\frac{1}{2} + 12\cdot\frac{1}{3} - 24\cdot\frac{1}{4}\right)\lambda +$$
$$+ \left(-90\cdot\frac{1}{2} - 12\cdot\frac{1}{2} + 24\cdot\frac{11}{24}\right)\lambda^2 + \left(12\cdot\frac{1}{6} - 24\cdot\frac{1}{4}\right)\lambda^3 +$$
$$+ 24\cdot\frac{1}{24}\lambda^4 = -20 - 56\lambda - 40\lambda^2 - 4\lambda^3 + \lambda^4.$$

§ 24*. Interpolation des fonctions de deux variables

Soit la fonction

$$z = f(x, y)$$

donnée sur un système de points équidistants (x_i, y_i) $(i, j = 0, 1, 2, \ldots)$, avec

$$x_i = x_0 + ih, \quad y_i = y_0 + jk,$$

de plus

$$h = \Delta x_i = \text{const}; \quad k = \Delta y_j = \text{const}.$$

Pour abréger l'écriture introduisons les notations

$$z_{ij} = f(x_i, y_j).$$

Les valeurs de la fonction z peuvent être rangées dans un *tableau à double entrée* (tableau 56).

Tableau 56

Valeurs d'une fonction de deux variables

$y \backslash x$	x_0	x_1	x_2	
y_0	z_{00}	z_{10}	z_{20}	...
y_1	z_{01}	z_{11}	z_{21}	...
y_2	z_{02}	z_{12}	z_{22}	...
...	...	...	...	...

L'interpolation d'une fonction de deux variables

$$z = f(x, y),$$

c'est-à-dire le calcul de ses valeurs non tabulées peut se faire, de proche en proche, séparément par rapport à chaque variable x et y. Supposons, par exemple, qu'il soit nécessaire d'obtenir la valeur

$$\bar{z} = f(\bar{x}, \bar{y}).$$

L'interpolation des fonctions dûment choisies d'une variable x:

$$f_k(x) = f(x, y_k),$$

où $y_k \approx \bar{y}$, permet de trouver les valeurs $f_k(\bar{x})$. A cet effet on utilise les lignes correspondantes du tableau double. En considérant les valeurs obtenues $f_k(\bar{x}) = f(\bar{x}, y_k)$ comme les valeurs de la fonction $f(\bar{x}, y)$ d'une seule variable y, on obtient à l'aide d'une des formules d'interpolation la valeur cherchée $f(\bar{x}, \bar{y}) = \bar{z}$.

On peut opérer également dans l'ordre inverse.

Exemple. Les valeurs de la *fonction de posteffet*

$$f(x, y) = \int_{-\infty}^{+\infty} e^{-y^2z^2-z-xe^{-z}}dz$$

sont données par le tableau suivant (cf. Yanke et Emde, « Tables des fonctions »)

y \ x	0,4	0,7	1,0
0,00	2,500	1,429	1,000
0,05	2,487	1,419	0,995
0,10	2,456	1,400	0,981

Trouver f (0,5 ; 0,03).

Solution. Formons les tableaux 57a, 57b, 57c en utilisant les lignes du tableau double donné.

$y = 0$ — *Tableau 57a*

x	f	Δf	$\Delta^2 f$
0,4	2,500	−1,071	0,642
0,7	1,429	−0,429	
1,0	1,000		

$y = 0,05$ — *Tableau 57b*

x	f	Δf	$\Delta^2 f$
0,4	2,487	−1,068	0,644
0,7	1,419	−0,424	
1,0	0,995		

$y = 0,10$ — *Tableau 57c*

x	f	Δf	$\Delta^2 f$
0,4	2,456	−1,056	0,637
0,7	1,400	−0,419	
1,0	0,981		

Puisque pour ces tableaux

$$h = 0,7 - 0,4 = 0,3,$$

en posant $x_0 = 0,4$, on a :

$$q = \frac{x - x_0}{h} = \frac{0,5 - 0,4}{0,3} = \frac{1}{3}.$$

On en tire successivement en utilisant la première formule de Newton :

$$f_0 = f(0,5;0) = 2{,}500 - \frac{1}{3}\cdot 1{,}071 + \frac{\frac{1}{3}\left(-\frac{2}{3}\right)}{2}\cdot 0{,}642 = 2{,}072;$$

$$f_1 = f(0{,}5; 0{,}05) = 2{,}487 - \frac{1}{3}\cdot 1{,}068 - \frac{1}{9}\cdot 0{,}644 = 2{,}069;$$

$$f_2 = f(0{,}5;\ 0{,}10) = 2{,}456 - \frac{1}{3}\cdot 1{,}056 - \frac{1}{9}\cdot 0{,}637 = 2{,}033.$$

Formons le tableau des valeurs obtenues (tableau 58).

Tableau 58

y	f	Δf	$\Delta^2 f$
0	2,072	−0,003	−0,033
0,05	2,069	−0,036	
0,10	2,033		

En adoptant $k = 0{,}05 - 0 = 0{,}05$ et $y_0 = 0$, on obtient :

$$q' = \frac{0{,}03 - 0}{0{,}05} = \frac{3}{5}.$$

D'où

$$f(0{,}5;\ 0{,}03) = 2{,}072 - \frac{3}{5}\cdot 0{,}003 + \frac{\frac{3}{5}\cdot\left(-\frac{2}{5}\right)}{2}\cdot(-0{,}033) = 2{,}074.$$

§ 25*. Différences à deux variables d'ordres supérieurs

Pour la fonction $z = f(x, y)$ donnée par le tableau double $\{z_{ij}\}$ on peut calculer les *différences partielles*

$$\Delta_x z_{ij} = z_{i+1,\,j} - z_{ij} \quad \text{et} \quad \Delta_y z_{ij} = z_{i,\,j+1} - z_{ij}.$$

En reprenant ces opérations, on obtient des *différences à deux variables d'ordres supérieurs*

$$\Delta^{m+n} z_{ij} = \Delta^{m+n}_{x^m y^n} z_{ij} = \Delta^m_{x^m}(\Delta^n_{y^n} z_{ij}) = \Delta^n_{y^n}(\Delta^m_{x^m} z_{ij}),$$

où l'on a posé $\Delta^{0+0} z_{ij} = z_{ij}$. Par exemple,

$$\Delta^{1+2} z_{ij} = \Delta_x(\Delta^2_{yy} z_{ij}) = \Delta_x(z_{i,\,j+2} - 2z_{i,\,j+1} + z_{ij}) =$$
$$= (z_{i+1,\,j+2} - 2z_{i+1,\,j+1} + z_{i+1,\,j}) - (z_{i,\,j+2} - 2z_{i,\,j+1} + z_{ij}).$$

§ 26*. Formule de Newton pour une fonction de deux variables

En recourant aux différences d'une fonction de deux variables $z = f(x, y)$, on peut former un polynôme d'interpolation analogue à celui de Newton. Soit $P(x, y)$ un polynôme entier tel que

$$\Delta^{m+n}_{x^m y^n} P(x_0, y_0) = \Delta^{m+n} z_{00} \qquad (1)$$

$(m, n = 0, 1, 2, \ldots)$. Supposons que $P(x, y)$ soit développé par rapport aux puissances généralisées des différences $x - x_0$ et $y - y_0$, c'est-à-dire

$$P(x, y) = c_{00} + c_{10}(x - x_0) + c_{01}(y - y_0) + c_{20}(x - x_0) \times \times (x - x_1) + c_{11}(x - x_0)(y - y_0) + + c_{02}(y - y_0)(y - y_1) + \ldots \qquad (2)$$

Posons $x = x_0$ et $y = y_0$ pour avoir, en vertu de la condition (1),

$$P(x_0, y_0) = z_{00} = c_{00}.$$

Composons les différences premières du polynôme $P(x, y)$

$$\Delta_x P(x, y) = c_{10}h + 2c_{20}h(x - x_0) + c_{11}h(y - y_0) + \ldots$$

et

$$\Delta_y P(x, y) = c_{01}k + c_{11}k(x - x_0) + 2c_{02}k(y - y_0) + \ldots$$

Il en résulte en posant $x = x_0$ et $y = y_0$ et en vertu de la condition (1):

$$\Delta_x P(x_0, y_0) = \Delta^{1+0} z_{00} = c_{10}h$$

et

$$\Delta_y P(x_0, y_0) = \Delta^{0+1} z_{00} = c_{01}k,$$

c'est-à-dire

$$c_{10} = \frac{\Delta^{1+0} z_{00}}{h}, \quad c_{10} = \frac{\Delta^{0+1} z_{00}}{k}.$$

Ensuite, en calculant les différences secondes du polynôme $P(x, y)$, on trouve:

$$\Delta_{xx} P(x, y) = 2!c_{20}h^2 + \ldots,$$
$$\Delta_{xy} P(x, y) = c_{11}hk + \ldots,$$
$$\Delta_{yy} P(x, y) = 2!c_{02}k^2 + \ldots$$

D'où pour $x = x_0$ et $y = y_0$:

$$\Delta_{xx} P(x_0, y_0) = \Delta^{2+0} z_{00} = 2!c_{20}h^2,$$
$$\Delta_{xy} P(x_0, y_0) = \Delta^{1+1} z_{00} = c_{11}hk,$$
$$\Delta_{yy} P(x_0, y_0) = \Delta^{0+2} z_{00} = 2!c_{02}k^2,$$

et

$$c_{20} = \frac{1}{2!} \cdot \frac{\Delta^{2+0} z_{00}}{h^2}, \quad c_{11} = \frac{\Delta^{1+1} z_{00}}{hk}; \quad c_{02} = \frac{1}{2!} \cdot \frac{\Delta^{0+2} z_{00}}{k^2}.$$

On opère d'une façon analogue pour obtenir les coefficients ultérieurs de la décomposition (2). En portant les valeurs des coefficients obtenues dans la formule (2) on compose le polynôme d'interpolation d'une fonction de deux variables

$$P(x, y) = z_{00} + \left[\frac{\Delta^{1+0}z_{00}}{h}(x-x_0) + \frac{\Delta^{0+1}z_{00}}{k}(y-y_0) \right] + \\ + \frac{1}{2!}\left[\frac{\Delta^{2+0}z_{00}}{h^2}(x-x_0)^{[2]} + 2\cdot\frac{\Delta^{1+1}z_{00}}{hk}(x-x_0)(y-y_0) + \\ + \frac{\Delta^{0+2}z_{00}}{k^2}(y-y_0)^{[2]} \right] + \ldots \qquad (3)$$

Pour interpoler la fonction $f(x, y)$, on pose:

$$f(x, y) \approx P(x, y).$$

En introduisant les variables

$$\frac{x-x_0}{h} = p, \qquad \frac{y-y_0}{k} = q$$

on rend généralement les calculs plus commodes; il vient alors

$$\frac{x-x_1}{h} = p-1, \qquad \frac{y-y_1}{k} = q-1,$$

etc. Par suite, la formule (3) se met sous la forme

$$z \approx z_{00} + (p\Delta^{1+0}z_{00} + q\Delta^{0+1}z_{00}) + \\ + \frac{1}{2!}[p(p-1)\Delta^{2+0}z_{00} + 2pq\Delta^{1+1}z_{00} + q(q-1)\Delta^{0+2}z_{00}] + \ldots, \qquad (4)$$

avec

$$x = x_0 + ph, \qquad y = y_0 + qk.$$

Si l'on pose $p = 0$ ou $q = 0$, (4) devient une formule de Newton correspondante.

Exemple. Appliquer la formule (4) et calculer $f = f(0,5; 0,03)$ de la fonction $f(x, y)$ de l'exemple du § 24.

Solution. En posant $x_0 = 0,4$, $y_0 = 0$, formons les tableaux des différences premières de la fonction f (tableaux 59a et 59b).

Tableau 59a

	$\Delta^{1+0}f_{0j}$	$\Delta^{1+0}f_{1j}$
$j=0$	−1,071	−0,429
$j=1$	−1,068	−0,424
$j=2$	−1,056	−0,419

Tableau 59b

	$i=0$	$i=1$	$i=2$
$\Delta^{0+1}f_{i0}$	−0,013	−0,010	−0,005
$\Delta^{0+1}f_{i1}$	−0,031	−0,019	−0,014

On en tire les différences secondes

$$\Delta^{2+0}f_{00} = \Delta^{1+0}f_{10} - \Delta^{1+0}f_{00} = -0{,}429 - (-1{,}071) = 0{,}642;$$
$$\Delta^{1+1}f_{00} = \Delta^{1+0}f_{01} - \Delta^{1+0}f_{00} = -1{,}068 - (-1{,}071) = 0{,}003$$

ou

$$\Delta^{1+1}f_{00} = \Delta^{0+1}f_{10} - \Delta^{0+1}f_{00} = -0{,}010 - (-0{,}013) = 0{,}003;$$
$$\Delta^{0+2}f_{00} = \Delta^{0+1}f_{01} - \Delta^{0+1}f_{00} = -0{,}031 - (-0{,}013) = -0{,}018.$$

Puisque

$$p = \frac{x - x_0}{h} = \frac{1}{3}; \quad q = \frac{y - y_0}{k} = \frac{3}{5},$$

en appliquant la formule (4), on obtient :

$$f = 2{,}500 + \frac{1}{3}\cdot(-1{,}071) + \frac{3}{5}\cdot(-0{,}013) + \frac{1}{2}\left[\frac{1}{3}\cdot\left(-\frac{2}{3}\right)\cdot 0{,}642 + 2\cdot\frac{1}{3}\cdot\frac{3}{5}\cdot 0{,}003 + \frac{3}{5}\cdot\left(-\frac{2}{5}\right)(-0{,}018)\right] = 2{,}500 - 0{,}357 - 0{,}0078 - 0{,}0713 + 0{,}0006 + 0{,}0021 = 2{,}067.$$

En comparant avec le résultat $f = 2{,}074$ obtenu par la première méthode on voit que les chiffres des millièmes méritent peu de confiance.

BIBLIOGRAPHIE

1. *E. Whittaker, G. Robinson.* The calculus of observations. A treatise on numerical mathematics. Blackie and Son, Ltd., London and Glasgow, 4e éd., 1944.
2. *V. Gontcharov.* Théorie d'interpolation et d'approximation des fonctions. GTTI, Moscou-Léningrad, 1934, chapitre I, §§ 18 à 21.
3. *J. Scarborough.* Numerical Mathematical Analysis. John Hopkins, 2e éd., 1950.
4. *V. Bradis.* Théorie et pratique des calculs. Outchpedguiz, Moscou, 1935, chapitre IX.
5. *W. E. Milne.* Numerical calculus. Princeton University Press, Princeton, 1949, chapitres III, IV.
6. *E. Rémez.* Méthodes numériques générales de l'approximation de Tchébychev. Editions de l'Académie des Sciences de la R.S.S. d'Ukraine, 1957, partie I, chapitre I.
7. Travaux pratiques sur les calculateurs, appareils de calcul et outils de calcul. Sous la direction générale de N. Lednev, «Sovietskaïa nauka», Moscou, 1959, chapitre III.
8. *V. Faddéeva.* Méthodes numériques d'algèbre linéaire. Gostekhizdat, Moscou-Léningrad, 1950, chapitre III, § 27.

CHAPITRE XV

DÉRIVATION APPROCHÉE

§ 1. Position du problème

Il arrive souvent que pour résoudre des problèmes pratiques il faut calculer les dérivées d'ordres imposés d'une fonction $y = f(x)$ donnée par un tableau. Il se peut également que l'expression analytique compliquée de cette fonction rende difficile sa dérivation immédiate. Ce sont autant de cas où l'on recourt à la *dérivation approchée*.

Les formules de dérivation approchée se déduisent en remplaçant la fonction donnée $f(x)$ sur le segment concerné $[a, b]$ par une fonction d'interpolation $P(x)$ (le plus souvent par un polynôme) et en posant ensuite:

$$f'(x) = P'(x) \qquad (1)$$

pour

$$a \leqslant x \leqslant b.$$

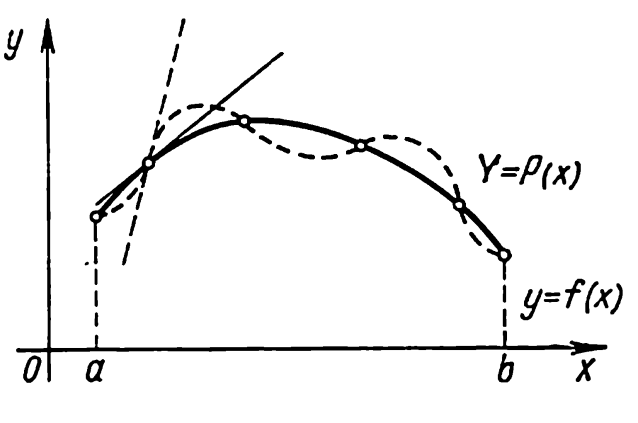


Fig. 65.

Les dérivées d'ordre supérieur de la fonction $f(x)$ s'obtiennent d'une façon analogue.

Si l'on connaît l'erreur

$$R(x) = f(x) - P(x)$$

de la fonction d'interpolation $P(x)$, l'erreur de la dérivée $P'(x)$ est donnée par la formule

$$r(x) = f'(x) - P'(x) = R'(x), \qquad (2)$$

c'est-à-dire *l'erreur de la dérivée d'une fonction d'interpolation est égale à la dérivée de l'erreur de cette fonction*. Il en est de même pour les dérivées d'ordre supérieur.

Il convient de noter que dans le cas général, la dérivation approchée est une opération moins précise que l'interpolation. En effet, le voisinage des ordonnées de deux courbes

$$y = f(x) \quad \text{et} \quad Y = P(x)$$

sur le segment $[a, b]$ ne garantit pas encore la proximité sur ce segment de leurs dérivées $f'(x)$ et $P'(x)$, c'est-à-dire un faible écart des coefficients angulaires des tangentes aux courbes considérées, les valeurs de l'argument étant les mêmes (fig. 65).

§ 2. Formules de dérivation approchée basées sur la première formule d'interpolation de Newton

Soit la fonction $y = f(x)$ donnée aux points équidistants x_i $(i = 0, 1, 2, \ldots, n)$ du segment $[a, b]$ par des valeurs $y_i = f(x_i)$. Pour chercher sur $[a, b]$ les dérivées $y' = f'(x)$, $y'' = f''(x)$, etc. *, remplaçons approximativement la fonction y par le polynôme d'interpolation de Newton établi pour un système de points $x_0, x_1, \ldots, x_k$ $(k \leqslant n)$.

On a

$$y(x) = y_0 + q\,\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0 + \frac{q(q-1)(q-2)(q-3)}{4!}\Delta^4 y_0 + \ldots, \quad (1)$$

où

$$q = \frac{x - x_0}{h} \quad \text{et} \quad h = x_{i+1} - x_i \quad (i = 0, 1, \ldots).$$

En multipliant les binômes, on obtient :

$$y(x) = y_0 + q\,\Delta y_0 + \frac{q^2 - q}{2}\Delta^2 y_0 + \frac{q^3 - 3q^2 + 2q}{6}\Delta^3 y_0 + \frac{q^4 - 6q^3 + 11q^2 - 6q}{24}\Delta^4 y_0 + \ldots \quad (1')$$

Puisque

$$\frac{dy}{dx} = \frac{dy}{dq}\cdot\frac{dq}{dx} = \frac{1}{h}\frac{dy}{dq},$$

il vient

$$y'(x) = \frac{1}{h}\left[\Delta y_0 + \frac{2q-1}{2}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{6}\Delta^3 y_0 + \frac{2q^3 - 9q^2 + 11q - 3}{12}\Delta^4 y_0 + \ldots\right]. \quad (2)$$

D'une façon analogue, comme

$$y''(x) = \frac{d(y')}{dx} = \frac{d(y')}{dq}\cdot\frac{dq}{dx},$$

---

* Bien entendu, le fait de l'existence des dérivées correspondantes de la fonction $f(x)$ doit être connu à l'avance, car s'il n'en est pas ainsi, les calculs ont un caractère illusoire.

on a

$$y''(x)=\frac{1}{h^2}\left[\Delta^2 y_0-(q-1)\Delta^3 y_0+\frac{6q^2-18q+11}{12}\Delta^4 y_0+\ldots\right]. \quad (3)$$

Quand la nécessité se présente, on procède de même pour calculer les dérivées de la fonction $y(x)$ d'un ordre quelconque.

Constatons que pour chercher les dérivées $y'(x)$, $y''(x)$, ... en un point fixé x, il faut prendre comme x_0 la valeur tabulée de l'argument la plus proche.

Quelquefois il faut chercher les dérivées de y aux points tabulaires principaux x_i. Dans ce cas les formules de dérivation numérique deviennent bien plus simples. Toute valeur tabulée pouvant être considérée comme initiale, posons, $x=x_0$, $q=0$; alors on a:

$$y'(x_0)=\frac{1}{h}\left(\Delta y_0-\frac{\Delta^2 y_0}{2}+\frac{\Delta^3 y_0}{3}-\frac{\Delta^4 y_0}{4}+\frac{\Delta^5 y_0}{5}-\ldots\right) \quad (4)$$

et

$$y''(x_0)=\frac{1}{h^2}\left(\Delta^2 y_0-\Delta^3 y_0+\frac{11}{12}\Delta^4 y_0-\frac{5}{6}\Delta^5 y_0+\ldots\right). \quad (5)$$

Si $P_k(x)$ est un polynôme de Newton qui contient les différences Δy_0, $\Delta^2 y_0$, ..., $\Delta^k y_0$ et si

$$R_k(x)=y(x)-P_k(x)$$

est l'erreur correspondante, l'erreur de la détermination de la dérivée s'écrit

$$R'_k(x)=y'(x)-P'_k(x).$$

On sait (chapitre XIV, § 15) que

$$R_k(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_k)}{(k+1)!}y^{(k+1)}(\xi)=$$
$$=h^{k+1}\frac{q(q-1)\ldots(q-k)}{(k+1)!}y^{(k+1)}(\xi),$$

où ξ est un certain nombre intermédiaire entre les valeurs x_0, x_1, ..., x_k et x. Aussi, en supposant que $y(x)\in C^{(k+2)}$ on obtient:

$$R'_k(x)=\frac{dR_k}{dq}\cdot\frac{dq}{dx}=\frac{h^k}{(k+1)!}\left\{y^{(k+1)}(\xi)\frac{d}{dq}[q(q-1)\ldots(q-k)]+\right.$$
$$\left.+q(q-1)\ldots(q-k)\frac{d}{dq}[y^{(k+1)}(\xi)]\right\}.$$

Supposant ensuite $\frac{d}{dq}[y^{(k+1)}(\xi)]$ bornée et en tenant compte du fait que $\frac{d}{dq}[q(q-1)\ldots(q-k)]_{q=0}=(-1)^k k!$, on en tire avec $x=x_0$ et, par suite, avec $q=0$,

$$R'_k(x_0)=(-1)^k\frac{h^k}{k+1}y^{(k+1)}(\xi). \quad (6)$$

Comme dans de nombreux cas $y^{(k+1)}(\xi)$ se prête mal à l'estimation, on pose approximativement pour un h petit:

$$y^{(k+1)}(\xi) \approx \frac{\Delta^{k+1} y_0}{h^{k+1}}$$

et donc

$$R'_k(x_0) \approx \frac{(-1)^k}{h} \frac{\Delta^{k+1} y_0}{k+1}. \tag{7}$$

D'une façon analogue on trouve l'erreur $R''_k(x_0)$ de la dérivée seconde $y''(x_0)$.

E x e m p l e 1. Chercher $y'(50)$ de la fonction $y = \lg x$ donnée par le tableau 60.

Tableau 60

Valeurs de la fonction $y = \lg x$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
50	1,6990	414	−36	5
55	1,7404	378	−31	
60	1,7782	347		
65	1,8129			

S o l u t i o n. Ici $h = 5$. Complétons le tableau 60 par les colonnes des différences finies (comme d'ordinaire, la place de la virgule n'est pas indiquée; elle est définie par les décimales des valeurs des fonctions).

En utilisant la première ligne du tableau on a, en vertu de la formule (4):

$$y'(50) = \frac{1}{5}(0{,}0414 + 0{,}0018 + 0{,}0002) = 0{,}0087.$$

Pour évaluer la précision de la valeur obtenue constatons que puisque la fonction tabulée ci-dessus est $y = \lg x$,

$$y'_x = \frac{M}{x} = \frac{0{,}43429}{x}.$$

Par conséquent,

$$y'(50) = \frac{0{,}43429}{50} = 0{,}0087.$$

Ainsi, les résultats coïncident à la quatrième décimale près.

E x e m p l e 2. La distance $y = f(t)$ parcourue par un point en mouvement rectiligne pendant le temps t est donnée par le tableau [1]:

i	Temps t_i en s	Distance $y(t_i)$ en cm	i	Temps t_i en s	Distance $y(t_i)$ en cm
0	0,00	0,000	5	0,05	35,721
1	0,01	1,519	6	0,06	50,000
2	0,02	6,031	7	0,07	65,798
3	0,03	13,397	8	0,08	82,635
4	0,04	23,396	9	0,09	100,000

Utiliser les différences jusqu'à l'ordre cinq y compris pour trouver la vitesse $V = \frac{dy}{dt}$ et l'accélération $W = \frac{d^2y}{dt^2}$ approchées du point aux instants $t = 0$; 0,01; 0,02; 0,03; 0,04.

Solution. Composons le tableau des différences (tableau 61).

Tableau 61

Différences de la fonction $y = f(t)$

i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$
0	1,519	2,993	−0,139	−0,082	−0,004
1	4,512	2,854	−0,221	−0,086	0,021
2	7,366	2,633	−0,307	−0,065	0,002
3	9,999	2,326	−0,372	−0,063	0,018
4	12,325	1,954	−0,435	−0,045	0,014
5	14,279	1,519	−0,480	−0,031	—
6	15,798	1,039	−0,511	—	
7	16,837	0,528	—		
8	17,365	—			
9	—				

Adoptant $h = 0{,}01$ et appliquant les formules (4) et (5) on obtient les valeurs approchées de la vitesse V (cm/s) et de l'accélération W (cm/s²). Par exemple,

$$V(0) = 100\,(1{,}519 - 1{,}496 - 0{,}046 + 0{,}020 - 0{,}001) = -0{,}4\ \text{cm/s},$$
$$W(0) = 10\,000\,(2{,}993 + 0{,}139 - 0{,}075 + 0{,}003) = 30600\ \text{cm/s}^2.$$

Les valeurs correspondantes de V et de W sont portées sur le tableau 62.

Remarquons que la loi du mouvement tabulée est donnée par la formule

$$y = 100\left(1 - \cos\frac{50\pi t}{9}\right).$$

Tableau 62

Valeurs de la vitesse V et de l'accélération W définies par la loi du mouvement $y = f(x)$

t	V	W	$\tilde{V}$	$\tilde{W}$
0,00	0,4	30 600	0,00	30 462
0,01	303,6	29 780	303,08	30 001
0,02	596,3	28 780	596,98	28 625
0,03	873,2	26 250	872,66	26 381
0,04	1121,7	23 360	1121,9	23 340

D'où

$$V = \frac{dy}{dt} = \frac{5000\pi}{9} \sin \frac{50\pi t}{9}$$

et

$$W = \frac{d^2y}{dt^2} = \frac{250000\pi^2}{81} \cos \frac{50\pi t}{9}.$$

Les deux colonnes droites du tableau 62 donnent à titre de comparaison les valeurs exactes $\tilde{V}$ et $\tilde{W}$.

Notons que les formules de dérivation approchée peuvent également être déduites de la deuxième formule de Newton.

§ 3. Formules de dérivation approchée basées sur la formule de Stirling

Les formules de dérivation numérique déduites au § 2 pour la fonction y au point $x = x_0$ ont l'inconvénient de n'utiliser que les valeurs unilatérales de la fonction pour $x > x_0$. Les formules de dérivation symétriques qui tiennent compte des valeurs de la fonction donnée y aussi bien pour $x > x_0$ que pour $x < x_0$ sont relativement plus exactes. Ces formules s'appellent en général *formules de dérivation par différences centrales.* Nous allons déduire l'une de ces formules en partant de la formule d'interpolation de Stirling.

Soient $\ldots, x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, \ldots$ un système de points équidistants à pas $x_{i+1} - x_i = h$ et $y_i = f(x_i)$ les valeurs correspondantes de la fonction donnée $y = f(x)$. Si l'on pose

$$q = \frac{x - x_0}{h}$$

et remplace approximativement la fonction y par le polynôme de Stirling, on aura:

$$y(x) = y_0 + q\Delta y_{-\frac{1}{2}} + \frac{q^2}{2!}\Delta^2 y_{-1} + \frac{q(q^2-1)}{3!}\Delta^3 y_{-\frac{3}{2}} + \frac{q^2(q^2-1)}{4!}\Delta^4 y_{-2} + \frac{q(q^2-1)(q^2-2^2)}{5!}\Delta^5 y_{-\frac{5}{2}} + \frac{q^2(q^2-1)(q^2-2^2)}{6!}\Delta^6 y_{-3} + \ldots, \quad (1)$$

où, pour abréger l'écriture, on introduit les notations

$$\Delta y_{-\frac{1}{2}} = \frac{\Delta y_{-1} + \Delta y_0}{2},$$

$$\Delta^3 y_{-\frac{3}{2}} = \frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2},$$

$$\Delta^5 y_{-\frac{5}{2}} = \frac{\Delta^5 y_{-3} + \Delta^5 y_{-2}}{2},$$

etc.

En tenant compte de

$$\frac{dq}{dx} = \frac{1}{h},$$

on obtient de la formule (1)

$$y'(x) = \frac{1}{h}\left(\Delta y_{-\frac{1}{2}} + q\Delta^2 y_{-1} + \frac{3q^2-1}{6}\Delta^3 y_{-\frac{3}{2}} + \frac{2q^3-q}{12}\Delta^4 y_{-2} + \right.$$
$$\left. + \frac{5q^4-15q^2+4}{120}\Delta^5 y_{-\frac{5}{2}} + \frac{3q^5-10q^3+4q}{360}\Delta^6 y_{-3} + \ldots\right), \quad (2)$$

$$y''(x) = \frac{1}{h^2}\left(\Delta^2 y_{-1} + q\Delta^3 y_{-\frac{3}{2}} + \frac{6q^2-1}{12}\Delta^4 y_{-2} + \right.$$
$$\left. + \frac{2q^3-3q}{12}\Delta^4 y_{-\frac{5}{2}} + \frac{15q^4-30q^2+4}{360}\Delta^6 y_{-3} + \ldots\right). \quad (2')$$

En particulier, si l'on pose $q = 0$, on a:

$$y'(x_0) = \frac{1}{h}\left(\Delta y_{-\frac{1}{2}} - \frac{1}{6}\Delta^3 y_{-\frac{3}{2}} + \frac{1}{30}\Delta^5 y_{-\frac{5}{2}} + \ldots\right) \quad (3)$$

et

$$y''(x_0) = \frac{1}{h^2}\left(\Delta^2 y_{-1} - \frac{1}{12}\Delta^4 y_{-2} + \frac{1}{90}\Delta^6 y_{-3} + \ldots\right). \quad (3')$$

Exemple 1. Calculer $y'(1)$ et $y''(1)$ de la fonction $y = y(x)$ donnée par le tableau 63.

Solution. En composant les différences de la fonction y (tableau 63) et en utilisant les termes soulignés on a en vertu de la formule (3):

$$y'(1) = \frac{1}{0,02}\left(-\frac{87\,355+88\,656}{2}\cdot 10^{-7} - \frac{1}{6}\cdot\frac{25+26}{2}\cdot 10^{-7} + \frac{1}{30}\cdot 1\cdot 10^{-7}\right) =$$
$$= -50\cdot(88\,005,5+4,2+0)\cdot 10^{-7} = -0,4400485.$$

Pour vérifier, constatons que la fonction tabulée est une fonction de Bessel à indice nul $y = J_0(x)$.

Tableau 63

Valeurs de la fonction $y = y(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0,96	0,7825361				
		−86029			
0,98	0,7739332		−1326		
		−87355		25	
1,00	0,7651977		−1301		1
		−88656		26	
1,02	0,7563321		−1275		
		−89931			
1,04	0,7473390				

On sait que

$$J_0'(1) = -J_1(x)\,|_{x=1} = -0{,}4400506.$$

D'une façon analogue, l'utilisation des termes soulignés d'un double trait et l'application de la formule (3′) amènent:

$$y''(1) = \frac{1}{0{,}02^2}\cdot\left(-1301\cdot10^{-7} - \frac{1}{12}\cdot 1\cdot 10^{-7}\right) =$$

$$= -2500\cdot1301\cdot10^{-7} \approx -3{,}2525\cdot10^{-1} - = 0{,}325250.$$

Pour comparer, voici la valeur exacte donnée par les relations entre les fonctions de Bessel

$$y''(1) = J_0''(1) = J_1(1) - J_0(1) =$$

$$= 0{,}4400506 - 0{,}7651977 = -0{,}325147.$$

Ainsi la recherche numérique de la dérivée seconde est en général une opération moins sûre que celle de la dérivée première.

R e m a r q u e. Parfois il faut trouver l'extrémum d'une fonction à dériver $y = y(x)$ donnée tabulairement. A cet effet, il faut que l'égalité $y'(\tilde{x}) = 0$ soit vraie au point de l'extrémum $\tilde{x}$. En annulant la dérivée $y'(x)$ de la formule (2) on trouve à l'aide de la méthode des approximations successives la valeur correspondante de q. On en tire

$$\tilde{x} = x_0 + qh,$$

et on calcule la valeur $\tilde{y}$ d'après la formule (1) ou l'une quelconque des formules d'interpolation. La valeur obtenue $\tilde{y}$ est un extrémum de la fonction si dans le voisinage du point $\tilde{x}$ le signe de la différence seconde $\Delta^2 y$ est constant.

Exemple 2. Trouver le zéro de la dérivée de la fonction $y = J_1(x)$ donnée par le tableau 64.

Tableau 64

Valeurs de la fonction $y = J_1(x)$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1,80	0,5815170			
		2561		
1,82	0,5817731		−1643	
		918		2
1,84	0,5818649		−1641	
		−723		4
1,86	0,5817926		−1637	
		−2360		2
1,88	0,5815566		−1635	
		−3995		
1,90	0,5811571			

Solution. Complétons le tableau 64 par les différences de la fonction y. Posons $x_0 = 1,84$. Utilisons les différences soulignées pour obtenir en vertu de la formule (2)

$$0 = \frac{918 - 723}{2} + q(-1641) + \frac{3q^2 - 1}{6} \cdot \frac{2+4}{2}$$

ou

$$0 = 97 - 1641q + \frac{3}{2} q^2.$$

Il en résulte que

$$q = \frac{97}{1641} + \frac{1}{1094} q^2. \tag{4}$$

Rejetons le petit terme non linéaire; on obtient alors la première approximation:

$$q^{(1)} = \frac{97}{1641} = 5,911 \cdot 10^{-2}.$$

En améliorant la précision de cette valeur, on obtient à partir de la formule (4) la deuxième approximation:

$$q^{(2)} = q^{(1)} + \frac{1}{1094} [q^{(1)}]^2 = 5,911 \cdot 10^{-2} + \frac{1}{1094} \cdot 3,494 \cdot 10^{-3} =$$
$$= 5,911 \cdot 10^{-2} + 3,2 \cdot 10^{-6} = 5,911 \cdot 10^{-2}.$$

Par conséquent, on peut poser:

$$q = 0,05911.$$

D'où

$$x = x_0 + qh = 1,84 + 0,05911 \cdot 0,02 = 1,8411822.$$

Ainsi

$$J_1'(1,8411822) = 0.$$

§ 4. Formules de dérivation numérique pour des points équidistants, exprimées par des valeurs de la fonction en ces points

Soient les points équidistants $x_0, x_1, x_2, \ldots, x_n$

$$x_{i+1} - x_i = h \qquad (i = 0, 1, 2, \ldots, n-1),$$

et soient les valeurs connues $y_i = y(x_i)$ $(i = 0, 1, \ldots, n)$ de la fonction $y = y(x)$. Formons pour le système donné des points x_i le polynôme d'interpolation de Lagrange (cf. chapitre XIV, § 12)

$$L_n(x) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(x)\, y_i}{(x - x_i)\, \Pi'_{n+1}(x_i)},$$

où

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n).$$

Il vient

$$L_n(x_i) = y_i \qquad (i = 0, 1, \ldots, n).$$

En posant

$$\frac{x - x_0}{h} = q,$$

on obtient

$$\Pi_{n+1}(x) = h^{n+1} q(q-1) \ldots (q-n) = h^{n+1} q^{[n+1]}$$

et

$$\Pi'_{n+1}(x_i) = (x_i - x_0)(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n) =$$
$$= h^n i(i-1) \ldots 1(-1) \ldots [-(n-i)] = (-1^{n-i}) h^n i!(n-i)! \quad (1)$$

Le polynôme de Lagrange $L_n(x)$ est donné donc par l'expression :

$$L_n(x) = \sum_{i=0}^{n} \frac{(-1)^{n-1} y_i}{i!(n-i)!} \cdot \frac{q^{[n+1]}}{q-i}. \qquad (2)$$

En retenant que

$$\frac{dx}{dq} = h,$$

on en tire

$$y'(x) \approx L'_n(x) = \frac{1}{h} \sum_{i=0}^{n} \frac{(-1)^{n-1} y_i}{i!(n-i)!} \frac{d}{dq} \left\{ \frac{q^{[n+1]}}{q-i} \right\}. \qquad (3)$$

D'une façon analogue on peut trouver les dérivées d'ordre supérieur de la fonction $y(x)$ donnée. Pour évaluer l'erreur

$$r_n(x) = y'(x) - L'_n(x)$$

faisons appel à la formule connue de l'erreur d'une formule d'interpolation (2) (chapitre XIV, § 14)

$$R_n(x) = y(x) - L_n(x) = \frac{y^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x), \tag{4}$$

où $\xi = \xi(x)$ est une valeur intermédiaire entre les points $x_0, x_1, \ldots$ $\ldots, x_n$ et x.

Supposons que $y(x) \in C^{(n+2)}$ pour déduire

$$r_n(x) = R_n'(x) = \frac{1}{(n+1)!}\left\{y^{(n+1)}(\xi)\,\Pi_{n+1}'(x) + \Pi_{n+1}(x)\frac{d}{dx}[y^{(n+1)}(\xi)]\right\}.$$

D'où l'on obtient, compte tenu de la formule (1) et supposant $\frac{d}{dx}[y^{(n+1)}(\xi)]$ bornée, l'erreur de la dérivée aux points

$$R_n'(x_i) = (-1)^{n-i}\,h^n\,\frac{i!\,(n-i)!}{(n+1)!}\,y^{(n+1)}(\xi), \tag{5}$$

ξ étant la valeur intermédiaire entre $x_0, x_1, \ldots, x_n$ et x.

I. Effectuons le calcul pour $n = 2$ (trois points). La formule (2) entraîne

$$L_2(x) = \frac{1}{2}y_0(q-1)(q-2) - y_1q(q-2) + \frac{1}{2}y_2q(q-1).$$

D'où, en tenant compte de ce que $\frac{dx}{dq} = h$, on aura:

$$y'(x) \approx L_2'(x) = \frac{1}{h}\left[\frac{1}{2}y_0(2q-3) - y_1(2q-2) + \frac{1}{2}y_2(2q-1)\right].$$

En particulier, pour les dérivées

$$y'(x_i) = y_i' \qquad (i = 0, 1, 2)$$

on obtient les expressions suivantes:

$$y_0' = \frac{1}{2h}(-3y_0 + 4y_1 - y_2);$$

$$y_1' = \frac{1}{2h}(-y_0 + y_2);$$

$$y_2' = \frac{1}{2h}(y_0 - 4y_1 + 3y_2)$$

aux erreurs respectives:

$$r_0 = \frac{1}{3}h^2y'''(\xi_0);$$

$$r_1 = -\frac{1}{6}h^2y'''(\xi_1);$$

$$r_2 = \frac{1}{3}h^2y'''(\xi_2).$$

Voici sans démonstration les formules de dérivation pour quatre et cinq points [3] que le lecteur peut facilement justifier lui-même.

II. $n = 3$ (quatre points):

$$y'_0 = \frac{1}{6h}(-11y_0 + 18y_1 - 9y_2 + 2y_3) - \frac{h^3}{4} y^{(4)}(\xi);$$
$$y'_1 = \frac{1}{6h}(-2y_0 - 3y_1 + 6y_2 - y_3) + \frac{h^3}{12} y^{(4)}(\xi);$$
$$y'_2 = \frac{1}{6h}(y_0 - 6y_1 + 3y_2 + 2y_3) - \frac{h^3}{12} y^{(4)}(\xi);$$
$$y'_3 = \frac{1}{6h}(-2y_0 + 9y_1 - 18y_2 + 11y_3) + \frac{h^3}{4} y^{(4)}(\xi).$$

III. $n=4$ (cinq points):

$$y'_0 = \frac{1}{12h}(-25y_0 + 48y_1 - 36y_2 + 16y_3 - 3y_4) + \frac{h^4}{5} y^{(5)}(\xi);$$
$$y'_1 = \frac{1}{12h}(-3y_0 - 10y_1 + 18y_2 - 6y_3 + y_4) - \frac{h^4}{20} y^{(5)}(\xi);$$
$$y'_2 = \frac{1}{12h}(y_0 - 8y_1 + 8y_3 - y_4) + \frac{h^4}{30} y^{(5)}(\xi);$$
$$y'_3 = \frac{1}{12h}(-y_0 + 6y_1 - 18y_2 + 10y_3 + 3y_4) - \frac{h^4}{20} y^{(5)}(\xi);$$
$$y'_4 = \frac{1}{12h}(3y_0 - 16y_1 + 36y_2 - 48y_3 + 25y_4) + \frac{h^4}{4} y^{(5)}(\xi).$$

L'examen des formules II et III montre que si le nombre de points est impair et si la dérivée est prise au milieu, la formule de dérivation numérique correspondante devient plus simple et est un peu plus exacte.

Fig. 66.

Ci-dessous nous donnons pour les cas $n = 2$ et $n = 4$ les formules de telles dérivées aux *différences centrales* [3]; pour rendre la symétrie évidente nous avons modifié la numérotation des points (fig. 66):

I. $n = 2$.

$$y'_0 = \frac{1}{2h}(y_1 - y_{-1}) - \frac{h^2}{6} y^{(3)}(\xi),$$

où $y_i = y(x_i)$ et $i = -1, 0, 1$;

II. $n=4$.

$$y'_0 = \frac{2}{3h}(y_1 - y_{-1}) - \frac{1}{12h}\left(y_2 - y_{-2}\right) + \frac{h^4}{30} y^{(5)}(\xi),$$

où $y_i = y(x_i)$ et $i = -2, -1, 0, 1, 2$.

§ 5. Dérivation graphique

Le problème de dérivation graphique consiste à construire d'après la courbe de la fonction $y = f(x)$ donnée la courbe de sa dérivée

$$Y = f'(x).$$

Soit la courbe de la fonction $y = f(x)$ (fig. 67). Pour construire à une échelle connue l la courbe de sa dérivée, on choisit sur cette courbe un réseau suffisamment serré de points 1, 2, 3, 4, 5, . . . qui comprend autant que possible les points remarquables du graphique. On mène à la levée par ces points avec le plus grand soin

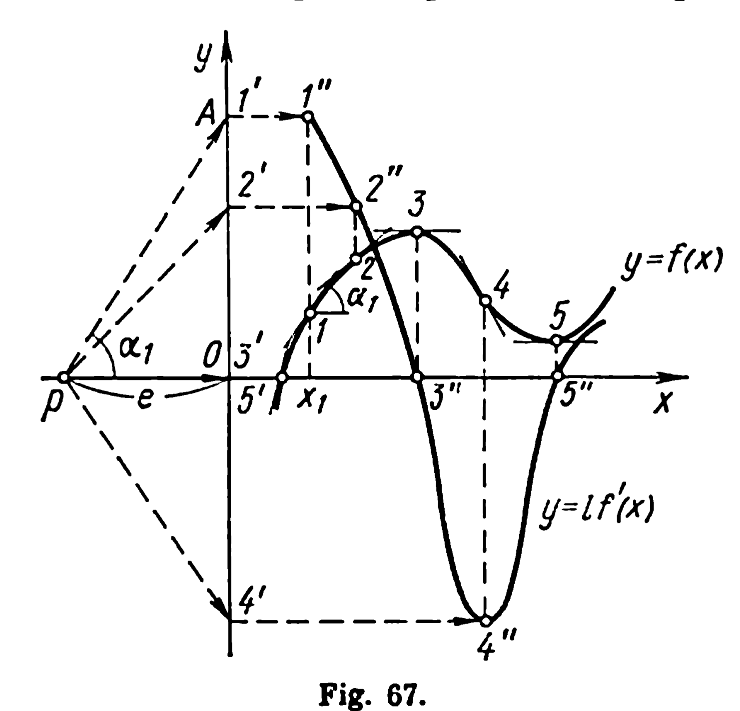


Fig. 67.

possible les tangentes à la courbe de la fonction. Ensuite, en choisissant sur l'axe Ox un point $P(-l, 0)$ (pôle) on mène les droites $P1'$, $P2'$, $P3'$, $P4'$, $P5'$, . . . parallèles aux tangentes respectives jusqu'à leur intersection avec l'axe Oy. Les segments de l'axe Oy: $01'$, $02'$, $03'$, $04'$, $05'$, . . . sont respectivement les grandeurs proportionnelles aux valeurs de la dérivée $y' = f'(x)$ aux points choisis, c'est-à-dire sont les ordonnées de la courbe de la dérivée. En effet, on a par exemple pour le point 1 de la figure 67:

$$OA = l \operatorname{tg} \alpha_1 = lf'(x_1).$$

Pour tous les autres points on obtient des résultats analogues. Les points d'intersection 1", 2", 3", 4", 5", . . . des parallèles menées par les points 1', 2', 3', 4', 5', . . . avec les verticales respectives qui passent par les points 1, 2, 3, 4, 5, . . . appartiennent donc à la courbe de la dérivée $y = lf'(x)$.

Si nous relions les points 1″, 2″, 3″, 4″, 5″, . . . par une ligne dont l'allure tient compte de la position des points intermédiaires, nous obtenons la courbe approchée de la dérivée y' à l'échelle l. En prenant $l = 1$, on obtient la courbe à l'échelle naturelle.

Pour que le graphique soit plus exact il est recommandé d'établir d'abord la direction de la tangente et de ne marquer qu'ensuite le point de tangence. A cette fin on divise la courbe de la fonction donnée en petits arcs qui diffèrent très peu d'un segment de droite. Considérons l'un de ces arcs AB (fig. 68). Construisons une famille de cordes parallèles à la sécante AB. Le lieu géométrique des milieux de ces cordes forme une courbe K qui coupe la courbe de la fonction en C, où la tangente est parallèle à la sécante AB. Ce procédé permet de déterminer sur chaque arc le point et la direction correspondante de la tangente. En poursuivant la construction on opère de la même façon.

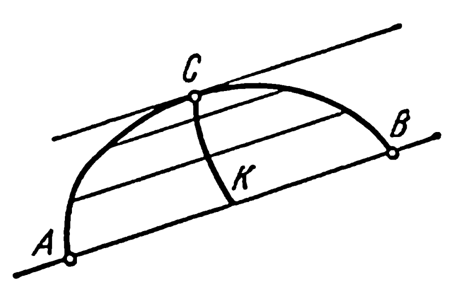


Fig. 68.

Pour plus de détails il faut se référer à des ouvrages spéciaux (cf. par exemple, [5]).

§ 6*. Notion de calcul approché des dérivées partielles

Si la fonction $z = f(x, y)$ est donnée par un réseau rectangulaire

$$x = x_0 + ih; \quad y = y_0 + jk$$

$(i, j = 0, 1, 2, \ldots)$, on peut la représenter approximativement par une formule d'interpolation (chapitre XIV, § 26)

$$\begin{aligned} z = z_{00} + [p\Delta^{1+0}z_{00} + q\Delta^{0+1}z_{00}] + \\ + \frac{1}{2!}\,[p(p-1)\,\Delta^{2+0}z_{00} + 2pq\Delta^{1+1}z_{00} + q(q-1)\,\Delta^{0+2}z_{00}] + \\ + \frac{1}{3!}\,[p(p-1)(p-2)\,\Delta^{3+0}z_{00} + 3p(p-1)\,q\Delta^{2+1}z_{00} + \\ + 3pq(q-1)\,\Delta^{1+2}z_{00} + q(q-1)(q-2)\,\Delta^{0+3}z_{00}] + \ldots, \end{aligned} \qquad (1)$$

où

$$p = \frac{x-x_0}{h}, \qquad q = \frac{y-y_0}{k}$$

et $\Delta^{m+n}z_{00} = \Delta^{m+n}_{x^m y^n} z\,(0, 0)$ sont des différences mixtes à deux variables.

La formule (1) conduit facilement aux dérivées partielles

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial p}\cdot\frac{dp}{dx} = \frac{1}{h}\,\frac{\partial z}{\partial p}, \quad \frac{\partial z}{\partial y} = \frac{\partial z}{\partial q}\cdot\frac{dq}{dy} = \frac{1}{k}\,\frac{\partial z}{\partial q},$$

etc.

BIBLIOGRAPHIE

1. *A. Krylov*. Conférences sur les calculs approchés, éd. 6, Gostekhizdat, Moscou, 1954, p. 228.
2. *J. B. Scarborough*. Numerical Mathematical Analysis. John Hopkins, 1950, chapitre VII.
3. *W. E. Milne*. Numerical calculus. Princeton University Press, Princeton, 1949, chapitre IV.
4. *Ch. Mikéladzé*. Méthodes numériques d'analyse mathématique. Gostekhizdat, Moscou, 1953, chapitre XII.
5. *C. Rungué*. Méthodes graphiques des calculs numériques. GTTI, Moscou-Léningrad, 1932, chapitre III, § 14.

CHAPITRE XVI

INTÉGRATION APPROCHÉE DES FONCTIONS

§ 1. Généralités

Si la fonction $f(x)$ est continue sur le segment $[a, b]$ et si l'on connaît sa primitive $F(x)$, l'intégrale définie de cette fonction dans les limites de a à b peut être calculée d'après la *formule de Newton-Leibniz*

$$\int_a^b f(x)\,dx = F(b) - F(a), \tag{1}$$

où $F'(x) = f(x)$.

Pourtant dans de nombreux cas la primitive $F(x)$ est trop compliquée ou ne peut s'obtenir à l'aide de procédés élémentaires; il en résulte que le calcul de l'intégrale définie d'après la formule (1) peut être trop difficile ou même pratiquement impossible.

Par ailleurs, dans la pratique, l'expression sous le signe somme $f(x)$ est donnée souvent tabulairement et la notion même de primitive perd alors tout son sens. Des problèmes analogues surgissent lors du calcul des intégrales multiples. C'est pourquoi les méthodes approchées et, en premier lieu, les *méthodes numériques* de calcul des intégrales définies acquièrent une grande importance.

Le problème de l'intégration numérique d'une fonction consiste à rechercher la valeur de l'intégrale définie à partir de plusieurs valeurs de la fonction sous le signe somme.

Le calcul numérique d'une intégrale simple s'appelle *quadrature mécanique*, celui d'une intégrale double, *cubature mécanique*. Les formules respectives sont dites *formules de quadrature* et *formules de cubature*.

Nous allons étudier d'abord le calcul numérique des intégrales simples. Le procédé usuel pour réaliser une quadrature consiste à remplacer la fonction donnée $f(x)$ sur le segment concerné $[a, b]$ par une fonction d'interpolation ou d'approximation $\varphi(x)$ simple (par un polynôme, par exemple), pour admettre approximativement ensuite

$$\int_a^b f(x)\,dx = \int_a^b \varphi(x)\,dx. \tag{2}$$

La fonction $\varphi(x)$ doit être telle que le calcul de l'intégrale $\int_a^b \varphi(x)\,dx$ soit immédiat.

Si la fonction $f(x)$ est donnée analytiquement, il faut évaluer l'erreur de la formule (2).

Considérons de plus près l'utilisation à cette fin du polynôme d'interpolation de Lagrange (chapitre XIV, § 12).

Supposons que pour la fonction $y = f(x)$ on connaît les valeurs correspondantes

$$f(x_i) = y_i \quad (i = 0, 1, 2, \ldots, n) \tag{3}$$

aux $n + 1$ points $x_0, x_1, x_2, \ldots, x_n$ du segment $[a, b]$. On demande de trouver approximativement

$$\int_a^b y\,dx = \int_a^b f(x)\,dx.$$

Formons le polynôme de Lagrange d'après les valeurs données y_i

$$L_n(x) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}\,y_i, \tag{4}$$

où

$$\Pi_{n+1}(x) = (x-x_0)(x-x_1)\ldots(x-x_n),$$

de plus,

$$L_n(x_i) = y_i \quad (i = 0, 1, 2, \ldots, n).$$

Remplaçons la fonction $f(x)$ par le polynôme $L_n(x)$ pour obtenir l'égalité

$$\int_a^b f(x)\,dx = \int_a^b L_n(x)\,dx + R_n[f], \tag{5}$$

où $R_n[f]$ est une erreur de la quadrature (5) (*reste*). L'application de l'expression (4) conduit à la formule de quadrature approchée

$$\int_a^b y\,dx = \sum_{i=0}^{n} A_i y_i, \tag{6}$$

avec

$$A_i = \int_a^b \frac{\Pi_{n+1}(x)}{(x-x_i)\,\Pi'_{n+1}(x_i)}\,dx \quad (i = 0, 1, 2, \ldots, n). \tag{7}$$

Si les limites d'intégration a et b sont des points d'interpolation, la formule de quadrature (6) est dite « formule du type fermé »; dans le cas contraire, on dit qu'elle est du « type ouvert ».

Pour calculer les coefficients A_i constatons que

1) les coefficients A_i pour la répartition donnée des points ne dépendent pas du choix de la fonction $f(x)$;

2) pour le polynôme de degré n, la formule (6) est exacte, puisque dans ce cas $L_n(x) \equiv f(x)$; par conséquent, en particulier, la formule (6) est exacte pour $y = x^k$ ($k = 0, 1, \ldots, n$), c'est-à-dire $R_n[x^k] = 0$ pour $k = 0, 1, \ldots, n$.

Si dans la formule (6) on pose $y = x^k$ ($k = 0, 1, 2, \ldots, n$), on obtient un système linéaire de $n + 1$ équations

$$\left.\begin{aligned} I_0 &= \sum_{i=0}^{n} A_i, \\ I_1 &= \sum_{i=0}^{n} A_i x_i, \\ &\cdots\cdots\cdots \\ I_n &= \sum_{i=0}^{n} A_i x_i^n, \end{aligned}\right\} \tag{8}$$

où

$$I_k = \int_a^b x^k\,dx = \frac{b^{k+1} - a^{k+1}}{k+1} \quad (k = 0, 1, \ldots, n),$$

qui permet de définir les coefficients $A_0, A_1, \ldots, A_n$ [1], [2]. Le déterminant du système (8) est un déterminant de Vandermonde

$$D = \prod_{i>j} (x_i - x_j) \neq 0.$$

Remarquons que l'application de cette méthode rend superflue la construction du polynôme de Lagrange $L_n(x)$.

S. Nikolski [3] a établi une méthode simple de calcul des erreurs des formules de quadrature.

E x e m p l e. Déduire une formule de quadrature de la forme

$$\int_0^1 y\,dx = A_0 y\left(\frac{1}{4}\right) + A_1 y\left(\frac{1}{2}\right) + A_2 y\left(\frac{3}{4}\right). \tag{9}$$

S o l u t i o n. Adoptons dans la formule (9)

$$y = x^k \quad (k = 0, 1, 2);$$

en tenant compte du fait que

$$\int_0^1 dx = 1, \quad \int_0^1 x\,dx = \frac{1}{2}, \quad \int_0^1 x^2\,dx = \frac{1}{3},$$

on obtient le système

$$\left.\begin{aligned} 1 &= A_0 + A_1 + A_2, \\ \frac{1}{2} &= \frac{1}{4} A_0 + \frac{1}{2} A_1 + \frac{3}{4} A_2, \\ \frac{1}{3} &= \frac{1}{16} A_0 + \frac{1}{4} A_1 + \frac{9}{16} A_2. \end{aligned}\right\}$$

D'où

$$A_0 = \frac{2}{3}, \quad A_1 = -\frac{1}{3}, \quad A_2 = \frac{2}{3}$$

et donc

$$\int_0^1 y\,dx = \frac{2}{3} y\left(\frac{1}{4}\right) - \frac{1}{3} y\left(\frac{1}{2}\right) + \frac{2}{3} y\left(\frac{3}{4}\right). \tag{10}$$

La formule de quadrature (10) du type ouvert est précisément la formule exacte de tous les polynômes de degré égal ou inférieur à deux. On voit facilement que pour $y = x^3$ la formule (10) donne également un résultat correct. Elle est donc exacte encore pour les polynômes de troisième degré.

§ 2. Formules de quadrature de Newton-Côtes

Supposons que pour la fonction donnée $y = f(x)$ il faille calculer l'intégrale

$$\int_a^b y\,dx.$$

Adoptons le pas

$$h = \frac{b-a}{n}$$

et découpons le segment $[a, b]$ à l'aide des points équidistants

$$x_0 = a, \quad x_i = x_0 + ih \quad (i = 1, 2, \ldots, n-1), \quad x_n = b$$

en n parties égales; soit

$$y_i = f(x_i) \quad (i = 0, 1, 2, \ldots, n).$$

Remplaçons la fonction y par le polynôme de Lagrange $L_n(x)$ correspondant pour obtenir la formule de quadrature approchée

$$\int_{x_0}^{x_n} y\,dx = \sum_{i=0}^{n} A_i y_i, \tag{1}$$

où A_i sont des constantes.

Déduisons les expressions explicites pour les constantes A_i de la formule (1).

On sait (chapitre XIV, § 12) que

$$L_n(x) = \sum_{i=0}^{n} p_i(x)\, y_i, \tag{2}$$

où

$$p_i(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}. \tag{3}$$

Introduisons les notations

$$q = \frac{x-x_0}{h} \tag{4}$$

et

$$q^{[n+1]} = q(q-1)\dots(q-n), \tag{5}$$

pour obtenir (cf. chapitre XV, § 4, formule (2)):

$$L_n(x) = \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!\,(n-i)!} \cdot \frac{q^{[n+1]}}{q-i}\, y_i. \tag{6}$$

En remplaçant dans la formule (1) la fonction y par le polynôme $L_n(x)$ on obtient, en vertu de la formule (6),

$$A_i = \int_{x_0}^{x_n} \frac{(-1)^{n-i}}{i!\,(n-i)!} \cdot \frac{q^{[n+1]}}{q-i}\, dx$$

ou, comme

$$q = \frac{x-x_0}{h}, \quad dq = \frac{dx}{h},$$

en faisant un changement de variables de l'intégrale définie, on amène

$$A_i = h \frac{(-1)^{n-i}}{i!\,(n-i)!} \int_0^n \frac{q^{[n+1]}}{q-i}\, dq \quad (i = 0, 1, 2, \dots, n).$$

Puisque

$$h = \frac{b-a}{n},$$

on pose ordinairement

$$A_i = (b-a)\, H_i,$$

où

$$H_i = \frac{1}{n} \frac{(-1)^{n-i}}{i!\,(n-i)!} \int_0^n \frac{q^{[n+1]}}{q-i}\, dq \quad (i = 0, 1, 2, \dots, n) \tag{7}$$

sont des constantes appelées *coefficients de Côtes* (cf., par exemple, [1], [4]).

La formule (1) se met alors sous la forme

$$\int_a^b y\,dx = (b-a)\sum_{i=0}^{n} H_i y_i, \tag{8}$$

avec

$$h=\frac{b-a}{n} \quad \text{et} \quad y_i = f_i(a+ih) \quad (i=0, 1, \ldots, n).$$

On voit sans peine que les relations

$$1)\ \sum_{i=0}^{n} H_i = 1\,; \qquad 2)\ H_i = H_{n-i}$$

sont vérifiées.

§ 3. Formule des trapèzes et son reste

En appliquant la formule (7) du paragraphe précédent pour $n = 1$, on a

$$H_0 = -\int_0^1 \frac{q(q-1)}{q}\,dq = \frac{1}{2},$$

$$H_1 = \int_0^1 q\,dq = \frac{1}{2}\,;$$

d'où

$$\int_{x_0}^{x_1} y\,dx = \frac{h}{2}(y_0+y_1). \tag{1}$$

Nous avons obtenu la *formule des trapèzes*, connue pour le calcul approché d'une intégrale définie (fig. 69).

Le reste (erreur) de la formule de quadrature (1) s'écrit

$$R = \int_{x_0}^{x_1} y\,dx - \frac{h}{2}(y_0+y_1).$$

Fig. 69.

Supposons que $y \in C^{(2)}[a, b]$ et déduisons une formule bien simple pour le calcul du reste. Considérons $R = R(h)$ comme une fonction du pas h; on peut alors poser

$$R(h) = \int_{x_0}^{x_0+h} y\,dx - \frac{h}{2}\left[y(x_0)+y(x_0+h)\right].$$

Dérivons cette formule deux fois par rapport à h

$$R'(h) = y(x_0+h) - \frac{1}{2}[y(x_0)+y(x_0+h)] - \frac{h}{2}y'(x_0+h) =$$

$$= \frac{1}{2}[y(x_0+h) - y(x_0)] - \frac{h}{2}y'(x_0+h)$$

et

$$R''(h) = \frac{1}{2}y'(x_0+h) - \frac{1}{2}y'(x_0+h) - \frac{h}{2}y''(x_0+h) = -\frac{h}{2}y''(x_0+h),$$

de plus

$$R(0) = 0, \quad R'(0) = 0.$$

En intégrant par rapport à h et en appliquant le théorème de la moyenne, on déduit successivement:

$$R'(h) = R'(0) + \int_0^h R''(t)\,dt = -\frac{1}{2}\int_0^h t y''(x_0+t)\,dt =$$

$$= -\frac{1}{2}y''(\xi_1)\int_0^h t\,dt = -\frac{h^2}{4}y''(\xi_1),$$

où $\xi_1 \in (x_0, x_0+h)$, et

$$R(h) = R(0) + \int_0^h R'(t)\,dt = -\frac{1}{4}\int_0^h t^2 y''(\xi_1)\,dt =$$

$$= -\frac{1}{4}y''(\xi)\int_0^h t^2\,dt = -\frac{h^3}{12}y''(\xi).$$

où $\xi \in (x_0, x_0+h)$.

Ainsi on a finalement:

$$R = -\frac{h^3}{12}y''(\xi), \qquad (2)$$

où $\xi \in (x_0, x_1)$.

Il s'ensuit en particulier que si $y'' > 0$, la formule (1) donne la valeur de l'intégrale *par excès*, et si $y'' < 0$, cette valeur est donnée *par défaut.*

§ 4. Formule de Simpson et son reste

La formule (7) du § 2 entraîne pour $n = 2$

$$H_0 = \frac{1}{2}\cdot\frac{1}{2}\int_0^2 (q-1)(q-2)\,dq = \frac{1}{4}\left(\frac{8}{3} - 6 + 4\right) = \frac{1}{6},$$

$$H_1 = -\frac{1}{2}\cdot\frac{1}{1}\int_0^2 q(q-2)\,dq = \frac{2}{3},$$

$$H_2 = \frac{1}{2}\cdot\frac{1}{2}\int_0^2 q(q-1)\,dq = \frac{1}{6}.$$

Donc, puisque $x_2 - x_0 = 2h$, on a :

$$\int_{x_0}^{x_2} y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2). \qquad (1)$$

La formule (1) s'appelle *formule de Simpson*. Son interprétation géométrique est donnée en remplaçant la courbe concernée $y = f(x)$ par une parabole $y = L_2(x)$ qui passe par trois points $M_0(x_0, y_0)$, $M_1(x_1, y_1)$ et $M_2(x_2, y_2)$ (fig. 70).

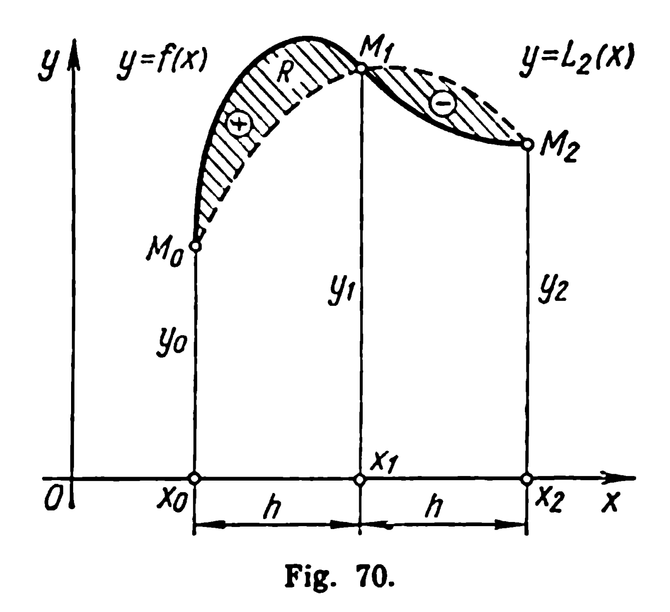


Fig. 70.

Le reste de la formule de Simpson s'écrit

$$R = \int_{x_0}^{x_2} y\,dx - \frac{h}{3}(y_0 + 4y_1 + y_2).$$

Supposant que $y \in C^{(4)}[a, b]$, on déduit d'une façon analogue à la formule des trapèzes une expression de R plus simple. En fixant le point médian x_1 et en considérant $R = R(h)$ comme une fonction du pas h $(h \geqslant 0)$, on obtient

$$R(h) = \int_{x_1-h}^{x_1+h} y\,dx - \frac{h}{3}[y(x_1 - h) + 4y(x_1) + y(x_1 + h)].$$

D'où, en dérivant trois fois de suite la fonction $R(h)$ par rapport à h, on aura

$$R'(h) = [y(x_1 + h) + y(x_1 - h)] - \frac{1}{3}[y(x_1 - h) + 4y(x_1) + y(x_1 + h)] -$$
$$- \frac{h}{3}[-y'(x_1 - h) + y'(x_1 + h)] = \frac{2}{3}[y(x_1 - h) + y(x_1 + h)] -$$
$$- \frac{4}{3}y(x_1) - \frac{h}{3}[-y'(x_1 - h) + y'(x_1 + h)];$$

$$R''(h) = \frac{2}{3}[-y'(x_1-h) + y'(x_1+h)] -$$

$$-\frac{1}{3}[-y'(x_1-h)+y'(x_1+h)] - \frac{h}{3}[y''(x_1-h)+y''(x_1+h)] =$$

$$= \frac{1}{3}[-y'(x_1-h)+y'(x_1+h)] - \frac{h}{3}[y''(x_1-h)+y''(x_1+h)];$$

$$R'''(h) = \frac{1}{3}[y''(x_1-h)+y''(x_1+h)] -$$

$$-\frac{1}{3}[y''(x_1-h)+y''(x_1+h)] - \frac{h}{3}[-y'''(x_1-h)+y'''(x_1+h)] =$$

$$= -\frac{h}{3}[y'''(x_1+h)-y'''(x_1-h)] = -\frac{2h^2}{3}y^{IV}(\xi_3),$$

où $\xi_3 \in (x_1-h,\ x_1+h)$.

En outre, on a:

$$R(0)=0,\quad R'(0)=0,\quad R''(0)=0.$$

Une intégration de proche en proche de $R'''(h)$ et l'application du théorème de la moyenne donnent

$$R''(h) = R''(0) + \int_0^h R'''(t)\,dt = -\frac{2}{3}\int_0^h t^2 y^{IV}(\xi_3)\,dt =$$

$$= -\frac{2}{3}y^{IV}(\xi_2)\int_0^h t^2\,dt = -\frac{2}{9}h^3 y^{IV}(\xi_2),$$

où $$\xi_2 \in (x_1-h,\ x_1+h);$$

$$R'(h) = R'(0) + \int_0^h R''(t)\,dt = -\frac{2}{9}\int_0^h t^3 y^{IV}(\xi_2)\,dt =$$

$$= -\frac{2}{9}y^{IV}(\xi_1)\int_0^h t^3\,dt = -\frac{1}{18}h^4 y^{IV}(\xi_1),$$

où $\xi_1 \in (x_1-h,\ x_1+h)$;

$$R(h) = R(0) + \int_0^h R'(t)\,dt = -\frac{1}{18}\int_0^h t^4 y^{IV}(\xi_1)\,dt =$$

$$= -\frac{1}{18}y^{IV}(\xi)\int_0^h t^4\,dt = -\frac{h^5}{90}y^{IV}(\xi),$$

où $\xi \in (x_1-h,\ x_1+h)$.

Ainsi le *reste de la formule de Simpson* vaut

$$R = -\frac{h^5}{90} y^{IV}(\xi), \tag{2}$$

où $\xi \in (x_0, x_2)$.

Cette formule est donc *exacte* non seulement pour les polynômes du deuxième, mais aussi du troisième degré, c'est-à-dire la formule de Simpson est plus exacte bien que le nombre d'ordonnées est relativement petit.

§ 5. Formules de Newton-Côtes d'ordres supérieurs

Les calculs correspondants pour $n = 3$ donnent suivant la formule (7) du § 2 la *formule de quadrature de Newton*

$$\int_{x_0}^{x_3} y\,dx = \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3) \tag{1}$$

(*règle des trois huitièmes*).

Le reste de la formule (1) est égal à [2]

$$R = -\frac{3h^5}{80} y^{IV}(\xi),$$

où $\xi \in (x_0, x_3)$.

Les formules de quadrature de Newton-Côtes d'ordres plus élevés sont données dans [1] et [2]. Les restes de ces formules sont établis par Steffensen (cf. [1], [5], [6]).

Remarquons que si la fonction $y = f(x)$ est suffisamment lisse, l'erreur de la formule de Newton-Côtes à $n + 1$ ordonnées est au moins de l'ordre [1], [6]

$$R = O\left[h^{2E\left(\frac{n}{2}\right)+3}\right],$$

$E\left(\frac{n}{2}\right)$ étant la partie entière de la fraction $\frac{n}{2}$.

Il s'ensuit qu'au sens de l'ordre de précision les formules de quadrature au nombre d'ordonnées impair sont plus avantageuses.

Tableau 65

Coefficients de Côtes

n	$\hat{H}_0$	$\hat{H}_1$	$\hat{H}_2$	$\hat{H}_3$	$\hat{H}_4$	$\hat{H}_5$	$\hat{H}_6$	$\hat{H}_7$	$\hat{H}_8$	Dénominateur commun N
1	1	1								2
2	1	4	1							6
3	1	3	3	1						8
4	7	32	12	32	7					90
5	19	75	50	50	75	19				288
6	41	216	27	272	27	216	41			840
7	751	3577	1323	2989	2989	1 323	3577	751		17 280
8	989	5888	−928	10 496	−4540	10 496	−928	5888	989	28 350

Voici à titre de référence le tableau des coefficients de Côtes (tableau 65). Pour la commodité de l'écriture, les coefficients de Côtes pour chaque n figurent sous la forme de fractions

$$H_i = \frac{\hat{H}_i}{N}$$

au dénominateur commun N. Pour vérifier constatons que

$$\sum_{i=0}^{n} \hat{H}_i = N.$$

Nous attirons l'attention sur le fait que pour de grands n les coefficients de Côtes peuvent être négatifs (cf., par exemple, $n = 8$).

Exemple. Calculer

$$I = \int_0^1 \frac{dx}{1+x},$$

en appliquant la formule de Newton-Côtes à sept coordonnées ($n = 6$).

Solution. Adoptant le pas

$$h = \frac{1-0}{6} = \frac{1}{6},$$

dressons le tableau 66 dans lequel on a posé pour commodité $\hat{H}_i = 840\, H_i$.

Tableau 66

Calcul de l'intégrale d'après la formule de Newton-Côtes

i	x_i	y_i	$\hat{H}_i$	$\hat{H}_i y_i$
0	0	1	41	41
1	$\frac{1}{6}$	$\frac{6}{7}$	216	185,142857
2	$\frac{1}{3}$	$\frac{3}{4}$	27	20,25
3	$\frac{1}{2}$	$\frac{2}{3}$	272	181,333333
4	$\frac{2}{3}$	$\frac{3}{5}$	27	16,2
5	$\frac{5}{6}$	$\frac{6}{11}$	216	117,818182
6	1	$\frac{1}{2}$	41	20,25
Σ				581,994372

D'où

$$I = \frac{1}{840} \cdot 581,994372 = 0,6933.$$

La valeur exacte est

$$I = \ln 2 = 0,69315 \ldots$$

Les coefficients de Côtes étant très compliqués dans le cas d'un grand nombre d'ordonnées, le calcul approché des intégrales définies s'opère pratiquement de la façon suivante : on divise l'intervalle d'intégration en un nombre suffisamment grand d'intervalles partiels pour appliquer à chacun de ces derniers la formule de quadrature de Newton-Côtes à petit nombre d'ordonnées (cf., par exemple, [7]). Les formules ainsi obtenues sont d'une structure plus simple et leur précision peut être aussi élevée que l'on veut.

Dans les paragraphes qui suivent nous considérerons des exemples de formules de ce type.

§ 6. Formule des trapèzes générale

Pour calculer l'intégrale

$$\int_a^b y\,dx$$

divisons l'intervalle d'intégration $[a, b]$ en n parties égales $[x_0, x_1]$, $[x_1, x_2]$, . . ., $[x_{n-1}, x_n]$ et appliquons à chacune d'elles la formule des trapèzes (cf. § 3 (1)). Posons $h = \frac{b-a}{n}$ et désignons par $y_i = f(x_i)$ $(i = 0, 1, \ldots, n)$ les valeurs de la fonction sous le signe somme aux points x_i ; il vient

$$\int_a^b y\,dx = \frac{h}{2}(y_0 + y_1) + \frac{h}{2}(y_1 + y_2) + \ldots + \frac{h}{2}(y_{n-1} + y_n)$$

ou

$$\int_a^b y\,dx = h\left(\frac{y_0}{2} + y_1 + y_2 + \ldots + y_{n-2} + y_{n-1} + \frac{y_n}{2}\right). \tag{1}$$

Géométriquement la formule (1) s'obtient en remplaçant la courbe de la fonction sous le signe somme $y = f(x)$ par une ligne brisée (fig. 71).

Si $y \in C^{(2)}[a, b]$, en vertu de (2) du § 3 le *reste* de la formule de quadrature (1) est égal à

$$R = \int_{x_0}^{x_n} y\,dx - \frac{h}{2}\sum_{i=1}^{n}(y_{i-1}+y_i) =$$

$$= \sum_{i=1}^{n}\left[\int_{x_{i-1}}^{x_i} y\,dx - \frac{h}{2}(y_{i-1}+y_i)\right] = -\frac{h^3}{12}\sum_{i=1}^{n} y''(\xi_i), \qquad (2)$$

où $\xi_i \in (x_{i-1}, x_i)$.

Considérons la moyenne arithmétique

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y''(\xi_i). \qquad (3)$$

Evidemment, μ est compris entre les valeurs minimale m_2 et maximale M_2 de la dérivée seconde y'' sur le segment $[a, b]$

$$m_2 \leqslant \mu \leqslant M_2.$$

Comme y'' est continue sur le segment $[a, b]$, elle prend comme valeurs sur $[a, b]$ tous les nombres intermédiaires entre m_2 et M_2.

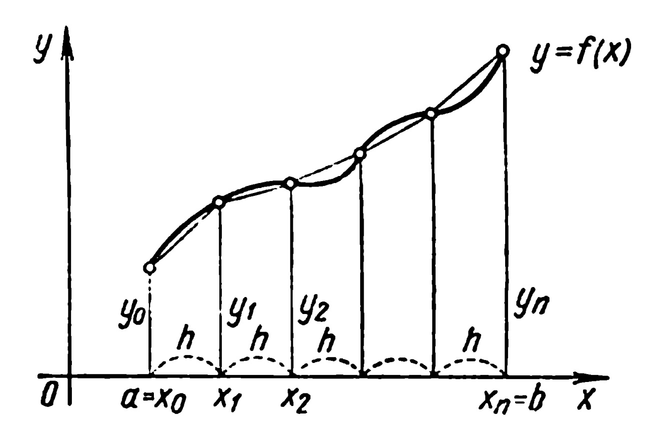


Fig. 71.

Il existe donc un point $\xi \in [a, b]$ tel que

$$\mu = f''(\xi).$$

Les formules (2) et (3) entraînent

$$R = -\frac{nh^3}{12}y''(\xi) = -\frac{(b-a)h^2}{12}y''(\xi),$$

où $\xi \in [a, b]$.

§ 7. Formule de Simpson générale (formule des paraboles)

Soient $n = 2m$ un nombre pair et $y_i = f(x_i)$ ($i = 0, 1, 2, \ldots, n$) les valeurs de la fonction $y = f(x)$ pour les points équidistants $a = x_0, x_1, \ldots, x_n = b$, dont le pas est

$$h = \frac{b-a}{n} = \frac{b-a}{2m}.$$

En appliquant la formule de Simpson (§ 4, (1)) à chaque intervalle

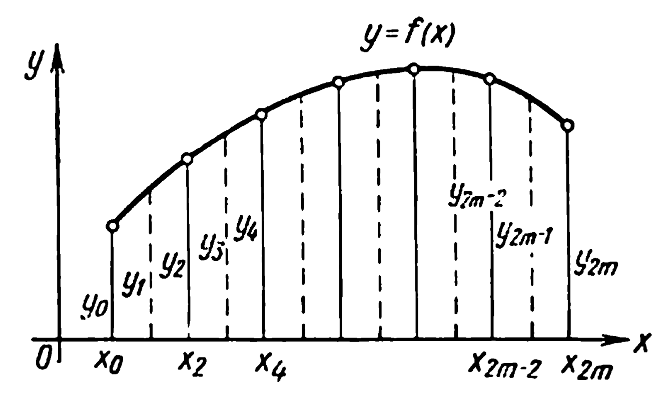


Fig. 72.

double $[x_0, x_2]$, $[x_2, x_4]$, ..., $[x_{2m-2}, x_{2m}]$ d'une longueur de $2h$ (fig. 72), on aura

$$\int_a^b y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2) + \frac{h}{3}(y_2 + 4y_3 + y_4) + \ldots \\ \ldots + \frac{h}{3}(y_{2m-2} + 4y_{2m-1} + y_{2m}).$$

On en tire la *formule de Simpson générale*

$$\int_a^b y\,dx = \frac{h}{3}[(y_0 + y_{2m}) + 4(y_1 + y_3 + \ldots + y_{2m-1}) + \\ + 2(y_2 + y_4 + \ldots + y_{2m-2})]. \quad (1)$$

Introduisons les notations

$$\sigma_1 = y_1 + y_3 + \ldots + y_{2m-1},$$
$$\sigma_2 = y_2 + y_4 + \ldots + y_{2m}$$

pour mettre la formule (1) sous une forme plus simple:

$$\int_a^b y\,dx = \frac{h}{3}[(y_0 + y_n) + 4\sigma_1 + 2\sigma_2]. \quad (1')$$

Si $y \in C^{(4)}[a, b]$, l'erreur de la formule de Simpson sur chaque intervalle double $[x_{2k-2}, x_{2k}]$ ($k = 1, 2, \ldots, m$) est donnée en

vertu du § 4, (2) par la formule

$$r_k = -\frac{h^5}{90} y^{\mathrm{IV}}(\xi_k),$$

où $\xi_k \in (x_{2k-2}, x_{2k})$. En additionnant toutes ces erreurs, on obtient le *reste de la formule de Simpson générale* sous la forme

$$R = -\frac{h^5}{90} \sum_{k=1}^{m} y^{\mathrm{IV}}(\xi_k).$$

Comme $y^{\mathrm{IV}}(x)$ est continue sur le segment $[a, b]$, il existe un point $\xi \in [a, b]$ tel que

$$y^{\mathrm{IV}}(\xi) = \frac{1}{m} \sum_{k=1}^{m} y^{\mathrm{IV}}(\xi_k).$$

On a donc

$$R = -\frac{mh^5}{90} y^{\mathrm{IV}}(\xi) = -\frac{(b-a)h^4}{180} y^{\mathrm{IV}}(\xi), \tag{2}$$

où $\xi \in [a, b]$.

Si l'on donne la borne d'erreur admissible $\varepsilon > 0$, en désignant

$$M_4 = \max |y^{\mathrm{IV}}(x)|,$$

le pas h sera déterminé par l'inégalité

$$(b-a)\frac{h^4}{180} M_4 < \varepsilon;$$

d'où

$$h < \sqrt[4]{\frac{180\varepsilon}{(b-a)M_4}},$$

c'est-à-dire l'ordre de h est $\sqrt[4]{\varepsilon}$.

Dans de nombreux cas il est malaisé d'évaluer d'après la formule (2) l'erreur de la formule de quadrature de Simpson (1). On applique alors le calcul double pour les pas h et $2h$ et on considère que les décimales qui coïncident sont celles de la valeur exacte de l'intégrale.

On peut indiquer encore un procédé pratiquement commode pour évaluer l'erreur de la formule de quadrature de Simpson. Supposons que sur le segment $[a, b]$ la dérivée $y^{\mathrm{IV}}(x)$ varie peu. En vertu de la formule (2) l'expression approchée de l'erreur cherchée est

$$R = Mh^4,$$

où le coefficient M est considéré comme constant. Soient Σ_h et Σ_H les valeurs approchées de l'intégrale

$$I = \int_a^b y\,dx,$$

fournies par la formule de Simpson respectivement pour les pas h et $H = 2h$. On a

$$I = \Sigma_h + Mh^4$$

et

$$I = \Sigma_H + M(2h)^4.$$

D'où

$$R = \frac{\Sigma_h - \Sigma_H}{15}.$$

Il est rationnel de prendre comme valeur approchée de l'intégrale I la valeur corrigée

$$I = \Sigma_h + \frac{\Sigma_h - \Sigma_H}{15}.$$

Constatons que si le nombre de divisions n est multiple de 4, le calcul de la somme Σ_H peut se faire à l'aide des valeurs tabulées, en les prenant deux à deux.

E x e m p l e. Calculer à l'aide de la formule de Simpson l'intégrale

$$I = \int_0^1 \frac{dx}{1+x},$$

en posant $n = 10$.

S o l u t i o n. On a $2m = 10$. D'où

$$h = \frac{1-0}{10} = 0,1.$$

Les résultats des calculs sont donnés dans le tableau 67.

Tableau 67

Calcul de l'intégrale suivant la formule de Simpson

i	x_i	y_{2j-1}	y_{2j}
0	0		$y_0 = 1,00000$
1	0,1	0,90909	
2	0,2		0,83333
3	0,3	0,76923	
4	0,4		0,71429
5	0,5	0,66667	
6	0,6		0,62500
7	0,7	0,58824	
8	0,8		0,55556
9	0,9	0,52632	
10	1,0		$0,50000 = y_n$
Σ		3,45955 (σ_1)	2,72818 (σ_2)

La formule (1′) donne

$$I \approx \frac{h}{3}(y_0 + y_n + 4\sigma_1 + 2\sigma_2) = 0{,}69315. \qquad (3)$$

Calculons l'erreur du résultat (3). L'erreur totale R se compose de l'erreur générée R_1 et du reste R_2. Il est clair que

$$R_1 = \sum_{i=0}^{n} A_i \varepsilon,$$

A_i étant les coefficients de la formule de Simpson et ε l'erreur d'arrondi maximale des valeurs de la fonction sous le signe somme.

Dans notre cas

$$R_1 = nh\varepsilon = (b - a)\,\varepsilon = 1 \cdot \frac{1}{2} \cdot 10^{-5} = 0{,}5 \cdot 10^{-5}.$$

Le reste est évalué d'après la formule (2). Puisque

$$y = \frac{1}{1+x} = (1+x)^{-1},$$

il vient

$$y^{\mathrm{IV}} = (-1)(-2)(-3)(-4)(1+x)^{-5} = \frac{24}{(1+x)^5}.$$

D'où

$$\max |y^{\mathrm{IV}}| = 24 \quad \text{pour} \quad 0 \leqslant x \leqslant 1$$

et donc

$$|R_2| \leqslant 1 \cdot \frac{(0{,}1)^4}{180} \cdot 24 = 1{,}3 \cdot 10^{-5}.$$

Ainsi, la borne d'erreur totale s'écrit

$$R = 0{,}5 \cdot 10^{-5} + 1{,}3 \cdot 10^{-5} = 1{,}8 \cdot 10^{-5} < 0{,}00002$$

et, par conséquent,

$$I = 0{,}69315 \pm 0{,}00002.$$

§ 8. Notion de la formule de quadrature de Tchébychev

Considérons la formule de quadrature

$$\int_{-1}^{1} f(t)\,dt = \sum_{i=1}^{n} B_i f(t_i), \qquad (1)$$

où B_i sont des constantes.

Tchébychev a proposé de choisir les abscisses t_i telles que

1) les constantes B_i soient égales entre elles;

2) la formule de quadrature (1) soit exacte pour tout polynôme jusqu'au degré n y compris.

Montrons comment dans ce cas on peut trouver les grandeurs B_i et t_i. En posant

$$B_1 = B_2 = \ldots = B_n = B$$

et en tenant compte de ce que pour $f(t) \equiv 1$ on a

$$2 = \sum_{i=1}^{n} B_i,$$

on obtient :

$$B = \frac{2}{n}.$$

La *formule de quadrature de Tchébychev* s'écrit donc

$$\int_{-1}^{1} f(t)\,dt = \frac{2}{n} \sum_{i=1}^{n} f(t_i). \tag{2}$$

Pour définir les abscisses t_i remarquons que d'après la condition 2) la formule (2) doit être exacte pour les fonctions de la forme

$$f(t) = t,\ t^2,\ \ldots,\ t^n.$$

Si l'on porte ces fonctions dans la formule (2), on obtient un système d'équations

$$\left.\begin{aligned} &t_1 + t_2 + \ldots + t_n = 0,\\ &t_1^2 + t_2^2 + \ldots + t_n^2 = \frac{n}{3},\\ &t_1^3 + t_2^3 + \ldots + t_n^3 = 0,\\ &t_1^4 + t_2^4 + \ldots + t_n^4 = \frac{n}{5},\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &t_1^n + t_2^n + \ldots + t_n^n = \frac{n\,[1-(-1)^{n+1}]}{2(n+1)}, \end{aligned}\right\} \tag{3}$$

qui permet de déterminer les inconnues t_i ($i = 1, 2, \ldots, n$). Tchébychev a montré que la résolution du système (3) se ramène à rechercher les racines d'une certaine équation algébrique de degré n [6], [8]. Le tableau 68 consigne les valeurs des solutions t_i du système (3) pour $n = 2, 3, \ldots, 7$.

Notons que le système (3), comme l'a montré S. Bernstein, ne possède pas pour $n = 8$ et $n \geqslant 10$ de solutions réelles. C'est là un inconvénient de principe de la formule de Tchébychev.

E x e m p l e 1. Déduire la formule de Tchébychev pour trois ordonnées ($n = 3$).

Tableau 68

Valeurs des abscisses t_i de la formule de Tchébychev

n	i	t_i	n	i	t_i
2	1; 2	∓0,577350	6	1; 6	∓0,866247
3	1; 3	∓0,707107		2; 5	∓0,422519
	2	0		3; 4	∓0,266635
4	1; 4	∓0,794654	7	1; 7	∓0,883862
	2; 3	∓0,187592		2; 6	∓0,529657
5	1; 5	∓0,832498		3; 5	∓0,323912
	2; 4	∓0,374541		4	0
	3	0			

S o l u t i o n. Pour déterminer les abscisses t_i $(i = 1, 2, 3)$ on a le système d'équations

$$\left.\begin{aligned} t_1+t_2+t_3&=0,\\ t_1^2+t_2^2+t_3^2&=1,\\ t_1^3+t_2^3+t_3^3&=0. \end{aligned}\right\} \tag{4}$$

Considérons les fonctions symétriques des solutions

$$C_1 = t_1 + t_2 + t_3,$$
$$C_2 = t_1t_2 + t_1t_3 + t_2t_3,$$
$$C_3 = t_1t_2t_3.$$

Le système (4) donne

$C_1 = 0$;

$$C_2 = \frac{1}{2}[(t_1+t_2+t_3)^2-(t_1^2+t_2^2+t_3^2)] = \frac{1}{2}(0-1) = -\frac{1}{2};$$

$$C_3 = \frac{1}{6}[(t_1+t_2+t_3)^3-3(t_1+t_2+t_3)(t_1^2+t_2^2+t_3^2)+2(t_1^3+t_2^3+t_3^3)] = \frac{1}{6}(0-0+0) = 0.$$

On en tire que les t_i sont les racines de l'équation auxiliaire

$$t^3 - C_1t^2 + C_2t - C_3 = 0$$

ou

$$t^3 - \frac{1}{2}t = 0.$$

On peut donc adopter

$$t_1 = -\frac{\sqrt{2}}{2}, \quad t_2 = 0, \quad t_3 = \frac{\sqrt{2}}{2}.$$

Ainsi la formule de Tchébychev correspondante s'écrit

$$\int_{-1}^{1} f(t)\,dt = \frac{2}{3}\left[f\left(-\frac{1}{\sqrt{2}}\right)+f(0)+f\left(\frac{1}{\sqrt{2}}\right)\right].$$

Pour appliquer la formule de quadrature de Tchébychev à l'intégrale de la forme

$$\int_{a}^{b} f(x)\,dx,$$

il faut la transformer en utilisant la substitution

$$x=\frac{b+a}{2}+\frac{b-a}{2}t,$$

qui associe le segment $a \leqslant x \leqslant b$ au segment $-1 \leqslant t \leqslant 1$. En appliquant la formule de Tchébychev (2) à l'intégrale transformée on aura

$$\int_{a}^{b} f(x)\,dx=\frac{b-a}{n}\sum_{i=1}^{n} f(x_i), \tag{5}$$

où

$$x_i=\frac{b+a}{2}+\frac{b-a}{2}t_i \tag{6}$$

et t_i ($i = 1, 2, \ldots, n$) forment la solution du système (3) (figurant dans le tableau 68).

La formule de quadrature de Tchébychev s'emploie essentiellement dans la construction navale.

E x e m p l e 2. Calculer l'intégrale

$$I=\int_{0}^{1}\frac{x\,dx}{1+x}$$

d'après la formule de Tchébychev à cinq ordonnées ($n = 5$).

S o l u t i o n. Introduisons les notations

$$f(x)=\frac{x}{1+x};$$

on a

$$I=\frac{1}{5}\left[f(x_1)+f(x_2)+f(x_3)+f(x_4)+f(x_5)\right],$$

où, en vertu de la formule (6),

$$x_1 = \frac{1}{2} + \frac{1}{2} t_1 = \frac{1}{2} + \frac{1}{2} \cdot (-0{,}83250) = 0{,}08375\,;$$

$$x_2 = \frac{1}{2} + \frac{1}{2} t_2 = \frac{1}{2} + \frac{1}{2} \cdot (-0{,}37454) = 0{,}31273\,;$$

$$x_3 = \frac{1}{2} + \frac{1}{2} t_3 = \frac{1}{2} + \frac{1}{2} \cdot 0 = 0{,}5\,;$$

$$x_4 = 1 - x_2 = 0{,}68727\,;$$

$$x_5 = 1 - x_1 = 0{,}91625.$$

Les valeurs correspondantes $y_i = f(x_i)$ $(i = 1, 2, 3, 4, 5)$ de la fonction sous le signe somme sont portées sur le tableau 69.

Tableau 69

Calcul de l'intégrale d'après la formule de Tchébychev

i	x_i	y_i
1	0,08375	0,0773
2	0,31273	0,2382
3	0,50000	0,3333
4	0,68727	0,4073
5	0,91625	0,4781
Σ		1,5342

D'où

$$I = \frac{1}{5} \cdot 1{,}5342 = 0{,}3068.$$

A titre de comparaison voici la valeur exacte de l'intégrale avec six décimales significatives

$$I = 0{,}306846\ldots$$

§ 9. Formule de quadrature de Gauss

Dans ce paragraphe nous appliquerons certains renseignements sur les polynômes de Legendre. On appelle polynômes de Legendre les expressions de la forme

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} [(x^2 - 1)^n] \quad (n = 0, 1, 2, \ldots).$$

Voici les propriétés fondamentales de ces polynômes [1]:

1) $P_n(1) = 1, \quad P_n(-1) = (-1)^n \quad (n = 0, 1, \ldots)$,

2) $\int\limits_{-1}^{1} P_n(x) \cdot Q_k(x)\, dx = 0 \quad (k < n)$, où $Q_k(x)$ est polynôme quelconque de degré k inférieur à n;

3) le polynôme de Legendre $P_n(x)$ possède n racines distinctes et réelles comprises dans l'intervalle $(-1, 1)$.

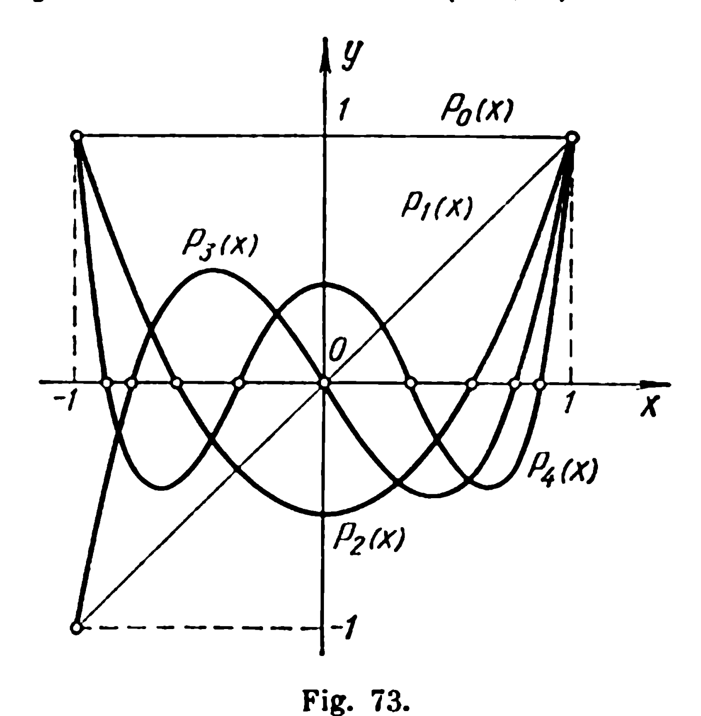


Fig. 73.

Ci-dessous nous donnons cinq polynômes de Legendre et leurs courbes (fig. 73):

$$P_0(x) = 1,$$

$$P_1(x) = x,$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1),$$

$$P_3(x) = \frac{1}{2}(5x^3 - 3x),$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

Déduisons maintenant la *formule de quadrature de Gauss*.

Considérons d'abord la fonction $y = f(t)$ définie sur le segment usuel $[-1; 1]$. Le cas général se ramène aisément à notre cas par substitution linéaire de la variable indépendante.

Voici la formulation du problème: comment sélectionner les points $t_1, t_2, \ldots, t_n$ et les coefficients $A_1, A_2, \ldots, A_n$ pour que la formule de quadrature

$$\int_{-1}^{1} f(t)\,dt = \sum_{i=1}^{n} A_i f(t_i) \tag{1}$$

soit exacte pour tout polynôme $f(t)$ de degré N le plus grand possible.

Puisque nous avons $2n$ constantes t_i et A_i ($i = 1, 2, \ldots, n$), alors que le polynôme de degré $2n - 1$ est défini par $2n$ coefficients, ce degré maximal dans le cas général est évidemment $N = 2n - 1$.

Pour garantir l'égalité (1) il faut et il suffit qu'elle soit vérifiée pour

$$f(t) = 1,\ t,\ t^2,\ \ldots,\ t^{2n-1}.$$

En effet, en posant

$$\int_{-1}^{1} t^k\,dt = \sum_{i=1}^{n} A_i t_i^k \qquad (k = 0, 1, 2, \ldots, 2n-1) \tag{2}$$

et

$$f(t) = \sum_{k=0}^{2n-1} C_k t^k,$$

on aura

$$\int_{-1}^{1} f(t)\,dt = \sum_{k=0}^{2n-1} C_k \int_{-1}^{1} t^k\,dt = \sum_{k=0}^{2n-1} C_k \sum_{i=1}^{n} A_i t_i^k =$$

$$= \sum_{i=1}^{n} A_i \sum_{k=0}^{2n-1} C_k t_i^k = \sum_{i=1}^{n} A_i f(t_i).$$

Ainsi, en tenant compte des relations

$$\int_{-1}^{1} t^k\,dt = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} \dfrac{2}{k+1} & \text{avec } k \text{ pair;} \\ 0 & \text{avec } k \text{ impair,} \end{cases}$$

on tire la conclusion que pour résoudre le problème posé [2], [3], [6], il suffit de déterminer t_i et A_i à partir du système de $2n$ équations

$$\left.\begin{aligned} &\sum_{i=1}^{n} A_i = 2, \\ &\sum_{i=1}^{n} A_i t_i = 0, \\ &\ldots\ldots\ldots\ldots \\ &\sum_{i=1}^{n} A_i t_i^{2n-2} = \frac{2}{2n-1}, \\ &\sum_{i=1}^{n} A_i t_i^{2n-1} = 0. \end{aligned}\right\} \tag{3}$$

Le système (3) est non linéaire et sa résolution par la voie usuelle présente de grandes difficultés. Toutefois, on peut appliquer à cet effet l'artifice suivant.

Considérons les polynômes

$$f(t) = t^k P_n(t) \qquad (k = 0, 1, \ldots, n-1),$$

où $P_n(t)$ est le polynôme de Legendre.

Les degrés de ce polynôme ne dépassant pas $2n-1$, ces polynômes doivent vérifier en vertu du système (3) la formule (1) et

$$\int_{-1}^{1} t^k P_n(t)\,dt = \sum_{i=1}^{n} A_i t_i^k P_n(t_i) \qquad (k=0, 1, \ldots, n-1). \tag{4}$$

D'autre part, l'orthogonalité des polynômes de Legendre (propriété (2)) rend vraies les égalités

$$\int_{-1}^{1} t^k P_n(t)\,dt = 0 \quad \text{pour} \quad k < n,$$

aussi

$$\sum_{i=1}^{n} A_i t_i^k P_n(t_i) = 0 \qquad (k=0, 1, \ldots, n-1). \tag{5}$$

Si l'on pose

$$P_n(t_i) = 0 \qquad (i = 1, 2, \ldots, n), \tag{6}$$

les égalités (5) seront nécessairement vraies quelles que soient les valeurs A_i. On sait (propriété (3)) que ces zéros sont réels, distincts et compris dans l'intervalle $(-1, 1)$. Si l'on connaît les abscisses t_i, on trouve facilement à partir du système linéaire des n premières équations du système (3) les constantes A_i $(i = 1, 2, \ldots, n)$. Le déterminant de ce sous-système est un déterminant de Vandermonde

$$D = \prod_{i>j} (t_i - t_j) \neq 0$$

et, par suite, la détermination des A_i est univoque.

On peut montrer que la formule (1) à coefficients ainsi déterminés est exacte pour tout polynôme de degré égal ou inférieur à $2n-1$.

La formule (1) où les t_i sont les zéros du polynôme de Legendre $P_n(t)$ et où les A_i $(i = 1, 2, \ldots, n)$ sont définies à partir du système (3) s'appelle *formule de quadrature de Gauss.*

E x e m p l e 1. Déduire la formule de Gauss pour le cas de trois ordonnées $(n = 3)$.

S o l u t i o n. Le polynôme de Legendre du troisième degré s'écrit

$$P_3(t)=\frac{1}{2}(5t^3-3t).$$

En annulant ce polynôme on obtient les racines

$$t_1=-\sqrt{\frac{3}{5}}\approx -0{,}774597\,;$$
$$t_2=0\,;$$
$$t_3=\sqrt{\frac{3}{5}}\approx 0{,}774597.$$

Pour déterminer les coefficients A_1, A_2, A_3 on a, en vertu de (3), le système

$$\left.\begin{aligned}A_1+A_2+A_3&=2,\\ -\sqrt{\frac{3}{5}}A_1+\sqrt{\frac{3}{5}}A_3&=0\,;\\ \frac{3}{5}A_1+\frac{3}{5}A_3&=\frac{2}{3},\end{aligned}\right\}$$

d'où

$$A_1=A_3=\frac{5}{9}\,,\qquad A_2=\frac{8}{9}.$$

Par conséquent,

$$\int_{-1}^{1} f(t)\,dt=\frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right)+8f(0)+5f\left(\sqrt{\frac{3}{5}}\right)\right].$$

Voici à titre de référence (tableau 70) les valeurs approchées des abscisses t_i et des constantes A_i de la formule de Gauss (1) pour $n=1$ à 8 (cf. [1], [4], [6]).

La formule de Gauss présente cet inconvénient que les abscisses des points t_i et les coefficients A_i sont en général des nombres irrationnels. Cet inconvénient est en partie compensé par une précision élevée en présence d'un nombre d'ordonnées relativement petit.

Examinons maintenant l'utilisation de la formule de Gauss pour calculer l'intégrale généralisée

$$\int_a^b f(x)\,dx.$$

En changeant la variable

$$x=\frac{b+a}{2}+\frac{b-a}{2}t,$$

Tableau 70

Eléments de la formule de Gauss

n	i	t_i	A_i
1	1	0	2
2	1; 2	∓0,57735027	1
3	1; 3 2	∓0,77459667 0	$\frac{5}{9}=0,55555556$ $\frac{8}{9}=0,88888889$
4	1; 4 2; 3	∓0,86113631 ∓0,33998104	0,34785484 0,65214516
5	1; 5 2; 4 3	∓0,90617985 ∓0,53846931 0	0,23692688 0,47862868 0,56888889
6	1; 6 2; 5 3; 4	∓0,93246951 ∓0,66120939 ∓0,23861919	0,17132450 0,36076158 0,46791394
7	1; 7 2; 6 3; 5 4	∓0,94910791 ∓0,74153119 ∓0,40584515 0	0,12948496 0,27970540 0,38183006 0,41795918
8	1; 8 2; 7 3; 6 4; 5	∓0,96028986 ∓0,79666648 ∓0,52553242 ∓0,18343464	0,10122854 0,22238104 0,31370664 0,36268378

on obtient

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b+a}{2}+\frac{b-a}{2}\,t\right)dt.$$

Appliquant à la dernière intégrale la formule de Gauss (1), on aura

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\sum_{i=1}^{n} A_i f(x_i), \tag{7}$$

où

$$x_i = \frac{b+a}{2} + \frac{b-a}{2} t_i \qquad (i = 1, 2, \ldots, n), \tag{8}$$

t_i étant les zéros du polynôme de Legendre $P_n(t)$, c'est-à-dire

$$P_n(t_i) = 0.$$

Le reste de la formule de Gauss (7) à n points est donné par l'expression [1], [6]:

$$R_n = \frac{(b-a)^{2n+1} (n\,!)^4 f^{(2n)}(\xi)}{[(2n)\,!]^3 (2n+1)},$$

d'où l'on tire

$$R_2 = \frac{1}{135} \left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi),$$

$$R_3 = \frac{1}{15750} \left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi),$$

$$R_4 = \frac{1}{3472875} \left(\frac{b-a}{2}\right)^9 f^{(8)}(\xi),$$

$$R_5 = \frac{1}{1237732650} \left(\frac{b-a}{2}\right)^{11} f^{(10)}(\xi),$$

$$R_6 = \frac{1}{648984486150} \left(\frac{b-a}{2}\right)^{13} f^{(12)}(\xi), \text{ etc.}$$

Exemple 2. Calculer l'intégrale

$$I = \int_0^1 \sqrt{1+2x}\, dx$$

en utilisant la formule de Causs à trois ordonnées ($n = 3$).

Solution. On a $a = 0$ et $b = 1$. En vertu de la formule (8) et du tableau 70 les abscisses des points avec cinq décimales significatives seront les suivantes:

$$x_1 = \frac{1}{2} + \frac{1}{2} t_1 = 0{,}11270;$$

$$x_2 = \frac{1}{2} + \frac{1}{2} t_2 = 0{,}50000;$$

$$x_3 = \frac{1}{2} + \frac{1}{2} t_3 = 0{,}88730.$$

Dans notre cas les coefficients respectifs de la formule (7) seront:

$$C_1 = \frac{b-a}{2} A_1 = \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{18} = 0{,}27778;$$

$$C_2 = \frac{b-a}{2} A_2 = \frac{1}{2} \cdot \frac{8}{9} = \frac{4}{9} = 0{,}44444;$$

$$C_3 = \frac{b-a}{2} A_3 = \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{18} = 0{,}27778.$$

Les calculs ultérieurs sont rangés dans le tableau 71.

39—01072

Tableau 71

Schéma de calcul d'une intégrale d'après la formule de Gauss

i	x_i	y_i	C_i	$C_i y_i$
1	0,11270	1,10698	0,27778	0,30747
2	0,50000	1,41421	0,44444	0,62853
3	0,88730	1,66571	0,27778	0,46270
Σ				1,39870

Par suite

$$I=\sum_{i=1}^{3} C_i y_i = 1{,}39870.$$

Pour évaluer le reste R_3, on peut utiliser la formule

$$R_3=\frac{1}{15750}\left(\frac{b-a}{2}\right)^7 f^{(6)}(\xi), \quad \text{où} \quad \xi\in(a, b).$$

Adoptant

$$f(x)=\sqrt{1+2x}=(1+2x)^{\frac{1}{2}},$$

on obtient

$$f^{(6)}(x)=\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)\left(-\frac{7}{2}\right)\left(-\frac{9}{2}\right)(1+2x)^{-\frac{11}{2}}\cdot 2^6=$$

$$=-945(1+2x)^{-\frac{11}{2}}.$$

D'où

$$\max|f^{(6)}(x)| = 945 \quad \text{pour} \quad 0\leqslant x\leqslant 1$$

et donc

$$|R_3|\leqslant\frac{945}{15750}\left(\frac{1}{2}\right)^7\approx\frac{1}{2000}.$$

Remarquons que la valeur exacte de l'intégrale est

$$I=\sqrt{3}-\frac{1}{3}\approx 1{,}39872.$$

§ 10. Certaines remarques sur la précision des formules de quadrature

Les formules de quadrature que nous venons d'étudier ont la structure suivante

$$\int_a^b f(x)\,dx=\sum_{i=1}^{n} A_i f(x_i)+R[f], \tag{1}$$

$x_1, x_2, \ldots, x_n$ étant le *système des points* donné appartenant au segment d'intégration $[a, b]$, A_i certaines *constantes* connues et $R[f]$ le reste.

Pour le même nombre d'ordonnées, la précision de différentes formules de quadrature n'est pas la même.

Exemple. Comparer la précision des formules différentes à trois ordonnées pour l'intégrale

$$I = \int_{-1}^{1} \sqrt{2+x}\,dx = 2\sqrt{3} - \frac{2}{3} = 2{,}797435\ldots$$

Solution. Appliquons la formule de Simpson pour obtenir

$$I \approx \frac{1}{3}\left[\sqrt{2-1} + 4\sqrt{2+0} + \sqrt{2+1}\right] = \frac{1}{3}\cdot 8{,}428905 = 2{,}809635.$$

La formule de Tchébychev donne le résultat suivant:

$$I \approx \frac{2}{3}\left[\sqrt{2-\frac{\sqrt{2}}{2}} + \sqrt{2+0} + \sqrt{2+\frac{\sqrt{2}}{2}}\right] = \frac{2}{3}\cdot 4{,}220097 = 2{,}813398.$$

Enfin la formule de Gauss fournit la valeur suivante:

$$I \approx 0{,}555566\left(\sqrt{2-0{,}774597} + \sqrt{2+0{,}774597}\right) + \\ + 0{,}888889\sqrt{2+0} = 2{,}797460.$$

Ainsi dans ce cas la formule de Gauss est la plus exacte.

Nous nous bornerons à l'étude des formules de quadrature *aux points équidistants*; ce sont en particulier les formules les plus usitées, celles des trapèzes, de Simpson, de Newton-Côtes. La précision de ces dernières est définie surtout par l'*ordre* du reste

$$R = O(h^m), \tag{2}$$

où

$$h = \frac{b-a}{n}$$

est le pas (n est le nombre de partitions) et m un nombre naturel. Par exemple, le reste de la formule des trapèzes s'écrit (§ 3):

$$R[f] = -\frac{b-a}{12} h^2 f''(\xi),$$

donc $m = 2$; celui de la formule de Simpson (§ 4):

$$R[f] = -\frac{b-a}{180} h^4 f^{IV}(\xi),$$

d'où $m = 4$. La précision de la formule de quadrature est considérée d'autant plus élevée que le nombre m est plus grand; dans ce sens la formule de Simpson est plus exacte que celle des trapè-

zes. La qualité de la formule se manifeste déjà avec un pas h suffisamment petit.

Il ne s'ensuit nullement que dans des cas concrets une formule plus grossière ne peut donner pour le même pas de meilleurs résultats qu'une formule exacte. Par exemple, pour la fonction (fig. 74)

$$f(x) = -8 + 45x^2 - 25x^4$$

on a

$$I = \int_{-1}^{1} f(x)\,dx = 2(-8+15-5) = 4.$$

Pour $h = 1$ la formule des trapèzes donne la valeur exacte

$$I_1 = \frac{1}{2} f(-1) + f(0) + \frac{1}{2} f(1) = 6 - 8 + 6 = 4$$

alors que la formule de Simpson pour $h = 1$ n'assure même pas le signe de l'intégrale

$$I_2 = \frac{1}{3}[f(-1) + 4f(0) + f(1)] =$$

$$= \frac{1}{3}(12 - 32 + 12) = -\frac{8}{3}.$$

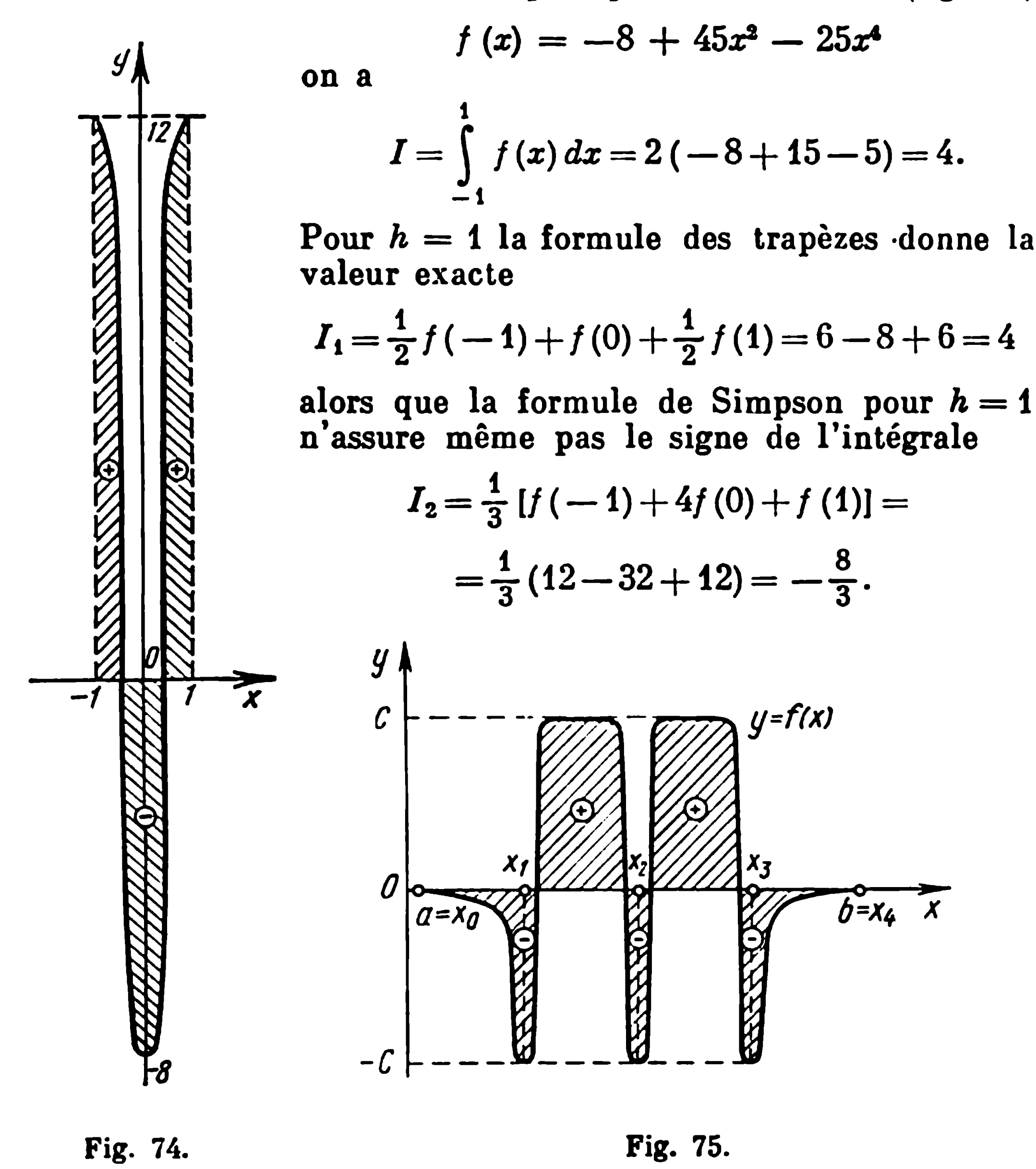


Fig. 74.

Fig. 75.

La précision d'une formule de quadrature pour un nombre de points fixe dépend sensiblement de la répartition de ces points. Si cette répartition est mauvaise, les résultats fournis par une formule de quadrature peuvent être très compromis. Par exemple, pour la fonction $y = f(x)$ représentée sur la figure 75, on obtient en choisissant les points équidistants $a = x_0$, x_1, x_2, x_3, $x_4 = b$ et

en utilisant la formule de Côtes correspondante à cinq ordonnées

$$I = \int_a^b f(x)\,dx < 0,$$

alors qu'il est clair que $I > 0$.

Il n'est pas difficile non plus de construire des exemples analogues pour une formule de quadrature quelconque à nombre d'ordonnées arbitraire.

En général, dans le cas d'un grand nombre de zéros de la fonction $f(x)$ sous le signe somme ou d'un grand nombre de ses extrémums (c'est-à-dire d'un grand nombre de zéros de la dérivée $f'(x)$) la précision des formules de quadrature diminue nettement par suite de grandes valeurs inévitables des dérivées supérieures. Aussi le pas h doit-il être choisi de sorte qu'il soit bien inférieur aux distances entre les zéros voisins de la fonction $f(x)$ et de sa dérivée $f'(x)$. A cette fin on recommande de partitionner le segment d'intégration principal $[a, b]$ en segments partiels $[\alpha, \beta]$ à l'intérieur desquels les fonctions $f(x)$ et $f'(x)$ restent du même signe (si c'est possible), et de calculer l'intégrale par parties en choisissant en général pour chaque segment partiel son propre pas. Dans des cas plus complexes il faut tenir compte également du comportement des dérivées d'ordre supérieur $f^{(n)}(x)$ $(n \geqslant 2)$. Pour une orientation générale, il convient de construire au préalable la courbe de la fonction sous le signe somme $y = f(x)$. Si cette fonction oscille fortement, il convient d'appliquer des procédés de calcul spéciaux. La précision des formules de quadrature peut être également améliorée par des procédés généraux établis à cet effet [9].

Lorsqu'on cherche la *borne d'erreur totale* d'une formule de quadrature (1), on doit également tenir compte de l'*erreur de sommation* R_1. Supposons que les termes de la somme $f(x_i)$ $(i = 1, 2, \ldots, n)$ sont calculés avec une erreur absolue égale ou inférieure à ε; et les coefficients A_i de la formule de quadrature sont des constantes positives exactes. On peut alors poser

$$R_1 \leqslant \sum_{i=1}^{n} A_i \varepsilon = \varepsilon \sum_{i=1}^{n} A_i. \tag{3}$$

La formule (1) étant vérifiée pour $f(x) \equiv 1$, il vient

$$\int_a^b dx = b - a = \sum_{i=1}^{n} A_i.$$

La formule (3) entraîne donc

$$R_1 \leqslant (b - a)\,\varepsilon. \tag{4}$$

Par conséquent, si l'on ne tient pas compte de l'erreur d'arrondi du résultat, la borne d'erreur totale de la formule de quadrature est

$$\bar{R} = (b - a)\,\varepsilon + |R[f]|,$$

où $|R[f]|$ est une *erreur de la méthode* qui peut être définie par le procédé indiqué dans ce qui précède.

Constatons que si la fonction $y = f(x)$ sous le signe somme est donnée par le tableau des valeurs $y_i = f(x_i)$ $(i = 1, 2, \ldots, n)$, alors en toute rigueur nous sommes dans l'impossibilité d'évaluer la précision de la formule de quadrature (1). Il en est ainsi parce que par un système fini de points $M_i(x_i, y_i)$ on peut mener un nombre illimité de courbes $y = f(x)$ (fig. 76) délimitant sur le segment donné $[a, b]$ des surfaces différentes, c'est-à-dire l'intégrale

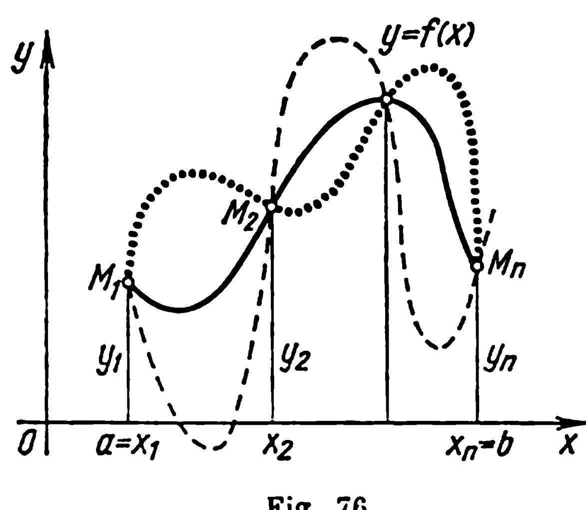


Fig. 76.

$$I = \int_a^b f(x)\,dx$$

peut avoir a priori une valeur parfaitement arbitraire (cf. fig. 76). L'application des formules de quadrature n'est alors admissible que dans le cas où l'on connaît dans une certaine mesure des valeurs intermédiaires non utilisées de la fonction sous le signe somme et ses propriétés générales qui permettent de juger sur l'allure de sa courbe.

§ 11*. Extrapolation suivant Richardson

Si l'on connaît l'ordre du reste $R = R[f]$ de la formule de quadrature (1) du § 10, la grandeur R peut être évaluée d'après la *méthode de calcul double*. Soit

$$R = O(h^m) \qquad (m \geqslant 1),$$

où

$$h = \frac{b-a}{n}$$

(n est le nombre de divisions); on peut poser approximativement

$$R = Mh^m, \tag{1}$$

où M est une certaine valeur considérée pour la fonction $f(x)$ sous le signe somme comme constante dans l'intervalle d'intégration

$[a, b]$. Choisissons deux pas distincts

$$h_1=\frac{b-a}{n_1} \quad \text{et} \quad h_2=\frac{b-a}{n_2},$$

où n_1 et n_2 ($n_2 > n_1$) sont les quantités de segments partiels dans le premier et le deuxième cas.

Désignons par I_{n_1} et I_{n_2} les valeurs approchées correspondantes de l'intégrale I. La formule (1) conduit à

$$R_{n_1}=I-I_{n_1}=M\left(\frac{b-a}{n_1}\right)^m \tag{2}$$

et

$$R_{n_2}=I-I_{n_2}=M\left(\frac{b-a}{n_2}\right)^m, \tag{2'}$$

R_{n_1} et R_{n_2} étant les restes correspondants. D'où

$$I_{n_2}-I_{n_1}=M(b-a)^m\left(\frac{1}{n_1^m}-\frac{1}{n_2^m}\right)$$

et

$$M=\frac{(n_1n_2)^m}{(b-a)^m}\cdot\frac{I_{n_2}-I_{n_1}}{n_2^m-n_1^m}$$

En vertu de la formule (1) l'expression du reste s'écrit

$$R=\left(\frac{n_1n_2}{n}\right)^m\cdot\frac{I_{n_2}-I_{n_1}}{n_2^m-n_1^m};$$

en particulier, pour $h=h_2$, c'est-à-dire pour $n=n_2$, on a:

$$R_{n_2}=\frac{n_1^m}{n_2^m-n_1^m}(I_{n_2}-I_{n_1}). \tag{3}$$

Utilisons la correction (3) et obtenons en vertu de la formule (2') pour l'intégrale I la valeur précisée:

$$I_{n_1,\,n_2}=I_{n_2}+\frac{n_1^m}{n_2^m-n_1^m}(I_{n_2}-I_{n_1}). \tag{4}$$

Ce procédé s'appelle *extrapolation suivant Richardson* [10]. Introduisons les notations

$$\frac{n_2}{n_1}=\alpha \quad (\alpha>1)$$

pour avoir

$$I_{n_1,\,n_2}=I_{n_2}+\beta(I_{n_2}-I_{n_1}), \tag{5}$$

où

$$\beta=\frac{1}{\alpha^m-1}. \tag{6}$$

Les coefficients β sont tabulés pour de différentes valeurs de α et m. Constatons que, pour la formule des trapèzes, $m=2$ et,

pour la formule de Simpson, $m = 4$. Le cas particulier de la formule (5) a été donné au § 7.

Montrons que si $I_{n_1} \neq I_{n_2}$, alors I_{n_1, n_2} se trouve toujours hors du segment $[I_{n_1}, I_{n_2}]$.

En effet, si

$$I_{n_2} > I_{n_1},$$

il suit de la formule (5) que

$$I_{n_1, n_2} > I_{n_2} = \max\{I_{n_1}, I_{n_2}\}.$$

Mais si

$$I_{n_2} < I_{n_1},$$

cette même formule (5) nous donne

$$I_{n_1, n_2} = I_{n_2} - \beta(I_{n_1} - I_{n_2}) < I_{n_2} = \min\{I_{n_1}, I_{n_2}\}.$$

Ainsi

$$I_{n_1, n_2} \bar{\in} [I_{n_1}, I_{n_2}],$$

c'est-à-dire I_{n_1, n_2} s'obtient de I_{n_1} et de I_{n_2} par *extrapolation*. D'où la dénomination de la méthode.

Tableau 72a

Extrapolation pour le cas de la formule des trapèzes

nºnº d'ordre		I_2	I_4	$I_{2,4}$	I
1	$I = \int_0^{\pi} \sin x \, dx$	1,571	1,896	2,004	2,000
2	$I = \int_0^{2} e^{-x^2} \, dx$	0,877	0,881	0,8823	0,8821
3	$I = \int_3^{7} x^2 \ln x \, dx$	185,7090	179,5385	177,4819	177,4836
4	$I = \int_0^{4} \frac{dx}{\sqrt{5-x^2}}$	0,9695	0,9389	0,9286	0,9267

nºnº d'ordre	$\varepsilon_1 = I - I_2$	$\varepsilon_2 = I - I_4$	$\varepsilon_{1,2} = I - I_{2,4}$
1	0,429	0,104	−0,004
2	0,0051	0,0011	−0,0002
3	−8,2254	−2,0549	0,0017
4	−0,0428	−0,0122	−0,0019

Tableau 72b

Extrapolation pour le cas de la formule de Simpson

nºnº d'ordre		I_2	I_4	$I_{2,4}$	I
1	$I=\int_0^{\pi} \sin x\, dx$	2,094	2,004	2,010	2,000
2	$I=\int_0^1 \frac{dx}{1+x^2}$	0,7833	0,7853	0,7855	0,7854
3	$I=\int_3^7 x^2 \ln x\, dx$	177,454	177,481	177,483	177,4836
4	$I=\int_0^4 \frac{dx}{(25-x^2)^{3/2}}$	0,0577	0,0541	0,0538	0,0533

nºnº d'ordre	$e_1=I-I_1$	$e_2=I-I_1$	$e_{1,2}=I-I_{2,4}$
1	−0,094	−0,004	−0,010
2	0,0021	0,0001	−0,0001
3	0,0296	0,0026	0,0006
4	−0,0044	−0,0008	−0,0005

Si $I_{n_1}=I_{n_2}$, il vient évidemment

$$I_{n_1,\,n_2}=I_{n_1}=I_{n_2}.$$

On peut montrer que pour une fonction $f(x)$ sous le signe somme suffisamment lisse, l'ordre du reste de $I_{n_1 . n_2}$ est au moins égal ou supérieur à $m+1$.

R e m a r q u e. Les tableaux 72a et 72b donnent des exemples d'extrapolation suivant Richardson.

Il apparaît de ces tableaux que pour les fonctions sans singularités l'extrapolation, en règle générale, améliore la précision des calculs.

On peut également déduire des formules d'extrapolation plus exactes en utilisant les valeurs I_{n_1}, I_{n_2} et I_{n_3} de l'intégrale cherchée, relatives à trois pas distincts

$$h_s=\frac{b-a}{ns} \quad (s=1,\ 2,\ 3)$$

et en tenant compte de deux premiers termes de la décomposition du reste de la formule de quadrature [10].

§ 12*. Nombres de Bernoulli

Considérons la fonction

$$f(x)=\frac{x}{e^x-1}. \tag{1}$$

En utilisant le développement connu

$$e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots,$$

on peut écrire:

$$f(x)=\frac{x}{\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots}=\frac{1}{1+\frac{x}{2!}+\frac{x^2}{3!}+\ldots}. \tag{2}$$

Il est donc évident que dans le voisinage de $x = 0$ la fonction $f(x)$ admet le développement en série entière qui, pour la commodité des calculs, peut être mise sous la forme

$$\frac{x}{e^x-1}=\sum_{n=0}^{\infty}\frac{B_n}{n!}x^n, \tag{3}$$

où $B_0 = f(0) = 1$. Pour déterminer les autres coefficients du développement B_n $(n = 1, 2, \ldots)$ qui s'appellent *nombres de Bernoulli* utilisons l'identité obtenue en vertu de la formule (2)

$$\sum_{n=0}^{\infty}\frac{x^n}{(n+1)!}\cdot\sum_{n=0}^{\infty}\frac{B_n}{n!}x^n\equiv 1.$$

En multipliant les séries entières entre elles et en annulant les coefficients des puissances positives de x on obtient un système infini d'équations linéaires

$$\frac{B_n}{n!}\cdot\frac{1}{1!}+\frac{B_{n-1}}{(n-1)!}\cdot\frac{1}{2!}+\ldots+\frac{B_0}{0!}\frac{1}{(n+1)!}=0 \quad (n=1, 2, 3, \ldots)$$

ou en multipliant par $(n+1)!$ et vu que

$$\frac{(n+1)!}{(n-k)!(k+1)!}=C_{n+1}^{n-k} \quad (k=0, 1, \ldots, n+1),$$

on aura

$$C_{n+1}^{1}B_n+C_{n+1}^{2}B_{n-1}+\ldots+C_{n+1}^{n}B_1+1=0. \tag{4}$$

Si l'on convient que

$$B_k=B^k, \tag{5}$$

la formule (4) peut se mettre sous la forme *symbolique abrégée*

$$(B+1)^{n+1}-B^{n+1}=0$$

ou, en remplaçant $n+1$ par n,

$$(B+1)^n - B^n = 0. \tag{6}$$

Posant $n=2, 3, 4, \ldots$ dans la formule (6), on obtient le système infini d'équations

$$\left.\begin{aligned} 2B_1+1&=0, \\ 3B_2+3B_1+1&=0, \\ 4B_3+6B_2+4B_1+1&=0, \\ 5B_4+10B_3+10B_2+5B_1+1&=0, \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}\right\} \tag{7}$$

On trouve successivement :

$$B_1=-\frac{1}{2};\quad B_2=\frac{1}{6};\quad B_3=0;\quad B_4=-\frac{1}{30};\quad B_5=0;$$

$$B_6=\frac{1}{42};\quad B_7=0;\quad B_8=-\frac{1}{30};\quad B_9=0;\quad B_{10}=\frac{5}{66};$$

$$B_{11}=0;\quad B_{12}=-\frac{691}{2730};\quad B_{13}=0;\quad B_{14}=\frac{7}{6};\quad B_{15}=0;$$

$$B_{16}=-\frac{3617}{510};\quad B_{17}=0;\quad B_{18}=\frac{43867}{798};\quad B_{19}=0;\quad B_{20}=-\frac{174611}{330},$$

etc.

Ainsi les nombres de Bernoulli peuvent être déterminés de proche en proche à partir de la formule symbolique (6); de plus, après le développement du binôme suivant la règle de Newton, les puissances du nombre B doivent être remplacées par les nombres de Bernoulli aux indices respectifs.

La fonction (1) s'appelle *fonction génératrice* des nombres de Bernoulli. En utilisant les notations (5), le développement (3) peut être mis sous la forme symbolique suivante:

$$\frac{x}{e^x-1}=e^{Bx}.$$

La structure du système (7) montre clairement que tous les nombres de Bernoulli sont rationnels. De plus, on a découvert que les nombres de Bernoulli aux indices impairs, sauf B_1, sont nuls. Démontrons cette propriété pour le cas général. Si l'on tient compte du fait que

$$B_0=1 \quad \text{et} \quad B_1=-\frac{1}{2},$$

on a

$$\varphi(x)=\frac{x}{e^x-1}-B_1x=\frac{x}{e^x-1}+\frac{x}{2}=1+\sum_{n=2}^{\infty}\frac{B_n}{n!}x^n. \tag{8}$$

Il est évident que

$$\varphi(x)=\frac{x(e^x+1)}{2(e^x-1)}=\frac{x}{2}\cdot\frac{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}{e^{\frac{x}{2}}-e^{-\frac{x}{2}}}=\frac{x}{2}\operatorname{cth}\frac{x}{2}$$

est une fonction paire. Son développement (8) ne contient donc que les puissances paires de x et, par conséquent,

$$B_n = 0 \quad \text{avec} \quad n = 3, 5, 7, \ldots$$

Les nombres de Bernoulli trouvent une application dans de nombreux domaines. En particulier, on les utilise dans la formule importante d'*Euler-Maclaurin* dont nous donnons ci-dessous la déduction.

§ 13*. Formule d'Euler-Maclaurin

Soit $y = f(x)$ une fonction définie dans le domaine $x \geqslant x_0$. Considérons l'opérateur de la *différence finie*

$$\Delta f(x) = f(x+h) - f(x),$$

où h est une valeur positive fixe. On entend naturellement par *opérateur inverse* $\frac{1}{\Delta}$ de la fonction $f(x)$ la fonction $F(x)$ qui vérifie l'équation aux différences finies

$$\Delta F(x) = f(x). \tag{1}$$

Ainsi l'équation (1) entraîne:

$$F(x)=\frac{1}{\Delta}f(x). \tag{2}$$

Si la fonction $f(x)$ est considérée sur un ensemble des points équidistants

$$x_0,\ x_1,\ x_2,\ \ldots,$$

où $\Delta x_i = x_{i+1} - x_i = h$ $(i = 0, 1, 2, \ldots)$, l'opérateur inverse $F(x_i) = \frac{1}{\Delta}f(x_i)$ se construit facilement. En effet, composons la somme finie

$$S(x_i)=\sum_{j=0}^{i-1} f(x_j) \quad (i=1, 2, \ldots),$$

en admettant par convention que $S(x_0) = 0$. On obtient évidemment

$$\Delta S(x_i) = S(x_{i+1}) - S(x_i) = f(x_i). \tag{3}$$

Par ailleurs, en vertu de l'équation (1), on a

$$\Delta F(x_i) = f(x_i). \tag{4}$$

Retranchant de l'égalité (4) l'égalité (3) on obtient:

$$\Delta [F(x_i) - S(x_i)] = 0$$

avec $i = 0, 1, 2, \ldots$ Par conséquent, la différence $F(x_i) - S(x_i)$ ne dépend pas de l'indice i et nous pouvons adopter:

$$F(x_i) - S(x_i) = F(x_0) - S(x_0) = F(x_0),$$

d'où

$$F(x_i) = F(x_0) + S(x_i),$$

$F(x_0)$ étant une grandeur constante arbitraire. Ainsi

$$\frac{1}{\Delta} f(x_i) = F(x_0) + S(x_i), \tag{5}$$

c'est-à-dire l'*opérateur inverse d'une différence finie est opérateur de sommation finie.*

Introduisons maintenant l'*opérateur de dérivation*

$$Df(x) = \frac{df(x)}{dx}.$$

Par *opérateur inverse* $\frac{1}{D}$ on entend l'opération d'intégration

$$\frac{1}{D} f(x) = \int_{x_0}^{x} f(x)\, dx.$$

En utilisant la série de Taylor, on trouve

$$\Delta f(x) = \sum_{k=1}^{\infty} \frac{h^k}{k!} D^k f(x) = \left\{ \sum_{k=1}^{\infty} \frac{h^k D^k}{k!} \right\} f(x) = (e^{hD} - 1) f(x).$$

Par conséquent,

$$\Delta = (e^{hD} - 1).$$

Il en résulte pour l'opérateur inverse $\frac{1}{\Delta}$ l'expression suivante:

$$\frac{1}{\Delta} = \frac{1}{e^{hD} - 1}.$$

En multipliant les deux membres de cette dernière égalité par hD, on aura:

$$hD \frac{1}{\Delta} = \frac{hD}{e^{hD} - 1}.$$

Ici le deuxième membre est la fonction génératrice des nombres de Bernoulli. Pour cette raison

$$hD \frac{1}{\Delta} = \sum_{k=0}^{\infty} \frac{B_k}{k!} h^k D$$

ou, avec plus de détail,

$$\frac{d}{dx}\left[\frac{1}{\Delta} f(x)\right]=\sum_{k=0}^{\infty}\frac{B_k}{k!}h^{k-1}D^k f(x). \tag{6}$$

En intégrant l'égalité (6) dans les limites de $x=x_0$ à $x=x_n$ et en utilisant la formule (5), on aura

$$\frac{1}{\Delta} f(x_n)-\frac{1}{\Delta} f(x_0)=\frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx+\sum_{k=1}^{\infty}\frac{B_k}{k!}h^{k-1}[f^{(k-1)}(x_n)-f^{(k-1)}(x_0)],$$

ou

$$F(x_0)+\sum_{j=0}^{n-1} f(x_j)-F(x_0)=\sum_{j=0}^{n-1} f(x_j)=$$

$$=\frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx+\sum_{k=1}^{\infty}\frac{B_k}{k!}h^{k-1}[f^{(k-1)}(x_n)-f^{(k-1)}(x_0)].$$

Vu que

$$B_1=-\frac{1}{2} \quad \text{et} \quad B_{2k+1}=0 \quad \text{pour} \quad k=1,\ 2,\ \ldots,$$

on obtient la *formule d'Euler-Maclaurin*

$$\int_{x_0}^{x_n} f(x)\,dx=h\left[\frac{1}{2} f(x_0)+f(x_1)+f(x_2)+\ldots+f(x_{n-1})+\frac{1}{2} f(x_n)\right]-$$

$$-\sum_{k=1}^{m}\frac{B_{2k}}{(2k)!}h^{2k}[f^{(2^k-1)}(x_n)-f^{(2^k-1)}(x_0)]+R_{2m}, \tag{7}$$

où R_{2m} est le *reste*. L'écriture de la formule (7) sous forme d'une série infinie n'est pas toujours légitime du fait qu'une série peut être divergente. En y portant les valeurs des nombres de Bernoulli, on aura

$$\int_{x_0}^{x_n} y\,dx=h\left(\frac{1}{2}y_0+y_1+y_2+\ldots+y_{n-1}+\frac{1}{2}y_n\right)-\frac{h^2}{12}(y'_n-y'_0)+$$

$$+\frac{h^4}{720}(y'''_n-y'''_0)-\frac{h^6}{30240}(y^{\mathrm{V}}_n-y^{\mathrm{V}}_0)+\ldots$$

$$\ldots-\frac{B_{2m}}{(2m)!}h^{2m}[f^{(2^m-1)}(x_n)-f^{(2^m-1)}(x_0)]+R_{2m}. \tag{8}$$

Le reste de la formule d'Euler-Maclaurin s'écrit [6]

$$R_{2m}=-nh^{2^m+3}\frac{B_{2m+2}}{(2m+2)!}f^{(2^m+2)}(\xi),$$

où $\xi\in(x_0,\ x_n)$.

La formule d'Euler-Maclaurin (8) s'emploie pour le calcul approché des intégrales définies, ainsi que pour la sommation approchée des valeurs des fonctions, les valeurs de l'argument étant équidistantes. En effet, il résulte de la formule (8) que

$$\sum_{i=0}^{n} f(x_i) = \frac{1}{h}\int_{x_0}^{x_n} f(x)\,dx + \frac{f(x_0)+f(x_n)}{2} +$$
$$+\sum_{k=1}^{m} \frac{B_{2k}}{(2k)!} h^{2^k-1} [f^{(2^k-1)}(x_n) - f^{(2^k-1)}(x_0)] + nh^{2^m+2} \frac{B_{2m+2}}{(2m+2)!} f^{(2^m+2)}(\xi). \tag{9}$$

Exemple 1. Utiliser la formule d'Euler-Maclaurin pour le calcul approché de l'intégrale définie

$$I = \int_{0,2}^{1} (\sin x - \ln x + e^x)\,dx.$$

Solution. Divisons le segment [0, 2; 1] en huit intervalles, par exemple, en prenant $h = 0,1$ et en posant

$$x_i = 0,2 + i\cdot 0,1 \quad (i = 0,\ 1,\ \ldots,\ 8).$$

Les résultats du calcul des valeurs correspondantes de la fonction $f(x) = \sin x - \ln x + e^x$ sont donnés dans le tableau 73.

Tableau 73

Valeurs de la fonction $f(x) = \sin x - \ln x + e^x$

x	0,2	0,3	0,4	0,5	0,6
$f(x)$	3,02951	2,84936	2,79754	2,82130	2,89759
x	0,7	0,8	0,9	1,0	
$f(x)$	3,01435	3,16605	3,34830	3,55975	

On en tire

$$\frac{1}{2} f(x_0) + f(x_1) + \ldots + f(x_7) + \frac{1}{2} f(x_8) = 24,1894.$$

En nous bornant à la dérivée d'ordre cinq, on aura:

$$f'(x)=\cos x-\frac{1}{x}+e^{x},$$

$$f'''(x)=-\cos x-\frac{2}{x^{3}}+e^{x},$$

$$f^{\mathrm{V}}(x)=\cos x-\frac{24}{x^{5}}+e^{x}.$$

Par suite

$f'(0,2)=-2,7985$; $f'(1)=2,2586$;

$f'''(0,2)=-249,7587$; $f'''(1)=0,1780$;

$f^{\mathrm{V}}(0,2)=-74\,997,7985$; $f^{\mathrm{V}}(1)=-20,7415$.

Portant les valeurs obtenues dans la formule (8), on obtient:

$$I=24,1894\cdot 0,1-\frac{(0,1)^{2}}{12}\cdot(2,2586+2,7985)+$$

$$+\frac{(0,1)^{4}}{720}\cdot(0,1780+249,7587)-\frac{(0,1)^{6}}{30\,240}\cdot(-20,7415+74997,7985)=$$

$$=2,41894-0,00421+0,00004=2,41477.$$

L'intégration immédiate donne

$$I=[-\cos x-x(\ln x-1)+e^{x}]|_{0,2}^{1}\approx 2,4148.$$

Exemple 2. Trouver la somme

$$\frac{1}{51^{2}}+\frac{1}{53^{2}}+\frac{1}{55^{2}}+\ldots+\frac{1}{99^{2}}.$$

Solution. Dans notre cas

$$f(x)=\frac{1}{x^{2}};\quad h=2;\quad x_{0}=51;\quad x_{n}=99.$$

Cherchons les dérivées d'ordre impair de la fonction $f(x)$:

$$f'(x)=-\frac{2}{x^{3}},$$

$$f'''(x)=-\frac{24}{x^{5}},$$

$$f^{\mathrm{V}}(x)=-\frac{720}{x^{7}},$$

$$f^{\mathrm{VII}}(x)=-\frac{40\,320}{x^{9}},\quad \text{etc.}$$

Portant ces valeurs dans la formule (9) et en nous bornant à la dérivée d'ordre sept, on aura:

$$\sum_{x=51}^{x=99} \frac{1}{x^2} = \frac{1}{2} \int_{51}^{99} \frac{dx}{x^2} + \frac{1}{2}\left(\frac{1}{51^2} + \frac{1}{99^2}\right) + \frac{1}{3}\left(\frac{1}{51^3} - \frac{1}{99^3}\right) -$$

$$-\frac{4}{15}\left(\frac{1}{51^5} - \frac{1}{99^5}\right) + \frac{16}{21}\left(\frac{1}{51^7} - \frac{1}{99^7}\right) - \frac{64}{15}\left(\frac{1}{51^9} - \frac{1}{99^9}\right) =$$

$$= 0{,}004\,753\,416 + 0{,}000\,243\,490 +$$
$$+ 0{,}000\,002\,169 - 0{,}000\,000\,001 = 0{,}004\,999\,074.$$

D'après la formule (9) dans laquelle on a posé $h = 2$, $n = 24$, $m = 4$, l'erreur du résultat obtenu est

$$R = 24 \cdot 2^{10} \cdot \frac{B_{10}}{8!} \cdot f^{(10)}(\xi) < 24 \cdot 2^{10} \cdot \frac{5}{66} \cdot \frac{1}{8!} \cdot \frac{11!}{50^{12}} < \frac{2}{25^{10}} \approx 10^{-14}.$$

§ 14. Calcul approché des intégrales impropres

L'intégrale

$$\int_a^b f(x)\,dx \tag{1}$$

s'appelle *propre* si

1) l'intervalle d'intégration $[a, b]$ est limité;

2) la fonction $f(x)$ sous le signe somme est continue sur $[a, b]$.

Dans le cas contraire l'intégrale (1) est dite *impropre*.

Considérons d'abord le calcul approché d'une intégrale impropre

$$\int_a^\infty f(x)\,dx \tag{2}$$

à *intervalle d'intégration illimité* où la fonction $f(x)$ est continue pour $a \leqslant x < \infty$.

L'intégrale (2) est dite *convergente* s'il existe une limite finie

$$\lim_{b\to\infty} \int_a^b f(x)\,dx, \tag{3}$$

et on pose par définition

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx. \tag{4}$$

Si la limite (3) n'existe pas, l'intégrale (2) est dite *divergente*, et on considère alors qu'elle n'a aucun sens. Avant de calculer une

intégrale impropre il faut donc s'assurer à l'aide des critères de convergence [10] que cette intégrale converge.

Pour calculer une intégrale impropre convergente (2) avec la précision imposée ε, mettons-la sous la forme

$$\int_a^\infty f(x)\,dx = \int_a^b f(x)\,dx + \int_b^\infty f(x)\,dx. \tag{5}$$

L'intégrale étant convergente, on peut choisir le nombre b suffisamment grand pour donner lieu à l'inégalité

$$\left|\int_b^\infty f(x)\,dx\right| < \frac{\varepsilon}{2}. \tag{6}$$

L'intégrale propre

$$\int_a^b f(x)\,dx$$

peut être calculée d'après l'une des formules de quadrature. Soit S la valeur approchée de cette intégrale à $\frac{\varepsilon}{2}$ près, c'est-à-dire

$$\left|\int_a^b f(x)\,dx - S\right| < \frac{\varepsilon}{2}. \tag{7}$$

Les formules (5), (6) et (7) entraînent

$$\left|\int_a^\infty f(x)\,dx - S\right| < \varepsilon,$$

le problème posé sera donc résolu.

Supposons maintenant que l'intervalle d'intégration $[a, b]$ soit limité et que la fonction $f(x)$ sous le signe somme ait un nombre fini de points de discontinuité sur $[a, b]$. Comme par hypothèse l'intervalle d'intégration peut être divisé en intervalles partiels avec un seul point de discontinuité de l'expression sous le signe somme, il suffit d'explorer le cas où sur $[a, b]$ la fonction $f(x)$ n'admet qu'un seul point de discontinuité c de deuxième espèce *.

* Si c est un point de discontinuité de première espèce, c'est-à-dire s'il existe des limites finies unilatérales

$$f(c-0) = \lim_{x\to c,\ x<c} f(x) \quad \text{et} \quad f(c+0) = \lim_{x\to c,\ x>c} f(x),$$

on peut alors poser

$$\int_a^b f(x)\,dx = \int_a^c f_1(x)\,dx + \int_c^b f_2(x)\,dx$$

Si c est un point intérieur du segment $[a, b]$, on adopte par définition

$$\int_a^b f(x)\,dx = \lim_{\substack{\delta_1 \to +0 \\ \delta_2 \to +0}} \left\{ \int_a^{c-\delta_1} f(x)\,dx + \int_{c+\delta_2}^b f(x)\,dx \right\} \tag{8}$$

et si cette limite existe, on dit que l'intégrale *converge*; dans le cas contraire on dit qu'elle *diverge*.

On définit de même la convergence de l'intégrale impropre (8) si le point de discontinuité c de la fonction $f(x)$ sous le signe somme coïncide avec l'une des extrémités de l'intervalle d'intégration $[a, b]$.

Pour calculer approximativement avec la précision imposée ε l'intégrale impropre convergente (8), où le point de discontinuité $c \in (a, b)$, on choisit les nombres positifs δ_1 et δ_2 tellement petits qu'ils donnent lieu à l'inégalité

$$\left| \int_{c-\delta_1}^{c+\delta_2} f(x)\,dx \right| < \frac{\varepsilon}{2}.$$

Ensuite, d'après les formules de quadrature connues on calcule approximativement les intégrales propres

$$\int_a^{c-\delta_1} f(x)\,dx \quad \text{et} \quad \int_{c+\delta_2}^b f(x)\,dx. \tag{9}$$

Il est évident que si S_1 et S_2 sont des valeurs approchées de l'intégrale (9) à $\frac{\varepsilon}{4}$ près, il vient

$$\int_a^b f(x)\,dx \approx S_1 + S_2$$

à ε près. Si le point de discontinuité c de la fonction sous le signe somme $f(x)$ est une extrémité de l'intervalle d'intégration $[a, b]$, la méthode de calcul change d'une façon manifeste.

---

où

$$f_1(x) = \begin{cases} f(x) & \text{si } a \leqslant x < c; \\ f(c-0) & \text{si } x = c; \end{cases} \quad \text{et}$$

$$f_2(x) = \begin{cases} f(c+0) & \text{si } x = c; \\ f(x) & \text{si } c < x \leqslant b; \end{cases}$$

de plus, les fonctions $f_1(x)$ et $f_2(x)$ sont continues respectivement sur les segments $[a, c]$ et $[c, b]$. Ainsi notre intégrale se ramène à la somme de deux intégrales propres.

§ 15. Méthode de L. Kantorovitch

Il arrive souvent que pour le calcul approché de l'intégrale d'une fonction discontinue, il soit utile d'appliquer la *méthode de Kantorovitch d'extraction des singularités* [1], [6], [10]. Cette méthode consiste en principe en la recherche d'une certaine fonction $g(x)$ possédant les mêmes singularités que la fonction $f(x)$, qui se prête à l'intégration élémentaire dans l'intervalle $[a, b]$ et telle que la différence $f(x) - g(x)$ soit suffisamment lisse sur le segment d'intégration $[a, b]$. Par exemple,

$$f(x) - g(x) \in C^{(m)}[a, b], \quad \text{où} \quad m \geqslant 1.$$

On aura alors

$$\int_a^b f(x)\,dx = \int_a^b g(x)\,dx + \int_a^b [f(x) - g(x)]\,dx,$$

où la première intégrale est prise directement, alors que la deuxième se calcule sans peine à l'aide des formules usuelles.

Considérons l'application de cette méthode au calcul de l'intégrale de la forme

$$\int_a^b \frac{\varphi(x)}{(x-x_0)^{\alpha}}\,dx, \tag{1}$$

où $x_0 \in [a, b]$, $0 < \alpha < 1$ et $\varphi(x)$ est continue sur le segment $[a, b]$.

Supposons que $\varphi(x) \in C^{(m+1)}[a, b]$, c'est-à-dire que $\varphi(x)$ possède sur le segment $[a, b]$ des dérivées continues jusqu'à l'ordre $(m+1)$ y compris.

Utilisant la formule de Taylor, on aura:

$$\varphi(x) = \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!}(x-x_0)^k + \psi(x), \tag{2}$$

où

$$\psi(x) = \varphi(x) - \sum_{k=0}^{m} \frac{\varphi^{k}(x_0)}{k!}(x-x_0)^k = \frac{\varphi^{(m+1)}(\xi)}{(m+1)!}(x-x_0)^{m+1} \tag{3}$$

$$(\xi \in (a, b)).$$

Il en résulte pour l'intégrale (1)

$$\int_a^b \frac{\varphi(x)\,dx}{(x-x_0)^{\alpha}} = \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!} \int_a^b (x-x_0)^{k-\alpha}\,dx + \int_a^b \frac{\psi(x)\,dx}{(x-x_0)^{\alpha}} =$$

$$= \sum_{k=0}^{m} \frac{\varphi^{(k)}(x_0)}{k!\,(k+1-\alpha)} [(b-x_0)^{k+1-\alpha} - (a-x_0)^{k+1-\alpha}] + I, \tag{4}$$

avec

$$I=\int_a^b \frac{\psi(x)\,dx}{(x-x_0)^{\alpha}}. \tag{5}$$

La formule (3) entraîne

$$\frac{\psi(x)}{(x-x_0)^{\alpha}} \in C^{(m)}[a,\ b]$$

(au moins!) ; par conséquent, l'intégrale (5) est une intégrale propre et peut être calculée avec une précision quelconque à l'aide d'une formule de quadrature convenable.

La méthode de Kantorovitch est également applicable aux intégrales impropres dont la fonction sous le signe somme possède plusieurs points de discontinuité de forme examinée. Dans ce cas, pour calculer l'intégrale, il suffit de diviser l'intervalle d'intégration en parties ne contenant qu'un point singulier de la fonction sous le signe somme et de mettre à profit l'additivité de l'intégrale.

Exemple 1. Calculer approximativement l'intégrale impropre [11]

$$I=\int_0^{\frac{1}{2}} \frac{dx}{\sqrt{x(1-x)}}.$$

Solution. La fonction sous le signe somme

$$f(x)=x^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}}$$

possède sur le segment $\left[0,\ \frac{1}{2}\right]$ un seul point singulier $x=0$. Développons la fonction

$$\varphi(x)=(1-x)^{-\frac{1}{2}}$$

en série de Taylor par rapport aux puissances de x, jusqu'à la puissance x^4. En appliquant le binôme de Newton, on aura:

$$\varphi(x)=1+\frac{1}{2}x+\frac{3}{8}x^2+\frac{5}{16}x^3+\frac{35}{128}x^4.$$

D'où

$$I=\int_0^{\frac{1}{2}} x^{-\frac{1}{2}}\,dx+\frac{1}{2}\int_0^{\frac{1}{2}} x^{\frac{1}{2}}\,dx+\frac{3}{8}\int_0^{\frac{1}{2}} x^{\frac{3}{2}}\,dx+\frac{5}{16}\int_0^{\frac{1}{2}} x^{\frac{5}{2}}\,dx+$$

$$+\frac{35}{128}\int_0^{\frac{1}{2}} x^{\frac{7}{2}}\,dx+I_1=\frac{715\,801}{645\,120}\sqrt{2}+I_1=1{,}5691585+I_1, \tag{6}$$

avec

$$I_1=\int_0^{\frac{1}{2}}\frac{\psi(x)}{\sqrt{x}}\,dx \tag{7}$$

et

$$\psi(x)=\frac{1}{\sqrt{1-x}}-\left(1+\frac{1}{2}x+\frac{3}{8}x^2+\frac{5}{16}x^3+\frac{35}{128}x^4\right);\quad \psi(0)=0.$$

L'intégrale propre (7) se calcule d'après la formule de Simpson, en posant $n = 10$ et le pas $h = \frac{1}{20} = 0{,}05$. Les résultats des calculs avec six décimales sont portés sur le tableau 74.

Tableau 74

Calcul de l'intégrale d'après la formule de Simpson

i	x_i	y_{2j-1}	y_{2j}
0	0		0,000000
1	0,05		
2	0,10	0,000000	0,000009
3	0,15		
4	0,20	0,000056	0,000216
5	0,25		
6	0,30	0,000624	0,001508
7	0,35		
8	0,40	0,003225	0,006316
9	0,45		
10	0,50	0,011588	0,020239
Σ		0,015493	0,008049

Il en résulte

$$I_1=\frac{1}{20\cdot 3}(0{,}020239+4\cdot 0{,}015493+2\cdot 0{,}008049)=$$

$$=\frac{1}{60}\cdot 0{,}098309=0{,}0016385.$$

Par conséquent, en vertu de la formule (6), on a

$$I=+\left.\begin{matrix}1{,}5691585\\0{,}0016385\end{matrix}\right\}=1{,}5707970.$$

Constatons que le calcul de l'intégrale I est élémentaire et que sa valeur exacte s'écrit

$$I=\frac{\pi}{2}=1,5707963\ \ldots$$

R e m a r q u e. Dans certains cas une intégrale impropre peut être transformée en une intégrale propre par substitution de la variable ou par intégration par parties.

E x e m p l e 2. Transformer en une intégrale propre l'intégrale

$$I=\int_{1}^{\infty}\frac{dx}{(1+x)\sqrt{x}}. \tag{8}$$

S o l u t i o n. En posant dans l'intégrale (8) $x=\frac{1}{z}$ on obtient une intégrale aux limites finies

$$I=\int_{0}^{1}\frac{dz}{(z+1)\sqrt{z}}=\int_{0}^{1}\frac{dx}{(1+x)\sqrt{x}} \tag{9}$$

qui devient singulière pour $x=0$.

En opérant une intégration par parties appropriée, on aura:

$$I=\int_{0}^{1}\frac{1}{1+x}\,d(2\sqrt{x})=\frac{2\sqrt{x}}{1+x}\Big|_{0}^{1}+\int_{0}^{1}2\sqrt{x}\,\frac{dx}{(1+x)^2}=1+2\int_{0}^{1}\frac{\sqrt{x}}{(1+x)^2}\,dx,$$

la dernière intégrale étant propre, l'application des formules de quadrature ne présente aucune difficulté.

§ 16. Intégration graphique

Le problème d'intégration graphique consiste à dresser d'après la courbe donnée d'une fonction continue $y=f(x)$ la courbe de sa primitive

$$F(x)=\int_{a}^{x}f(x)\,dx.$$

Autrement dit, il faut construire une courbe $y=F(x)$ telle qu'en tout point x de cette courbe l'ordonnée soit numériquement égale à l'aire du trapèze curviligne de base $[a, x]$, limitée par la courbe donnée $y=f(x)$.

Pour construire approximativement la courbe de la primitive $y=F(x)$ l'aire du trapèze curviligne correspondant limité par la courbe $y=f(x)$ est divisée en bandes verticales étroites à l'aide des parallèles à l'axe des y aux points $x_0, x_1, \ldots$ $(a=x_0<x_1<$ $<x_2<\ldots)$ (fig. 77). Appliquons le théorème de la moyenne pour

remplacer, s'il est possible, chacune de ces bandes par un rectangle de surface équivalente ayant la même base et la hauteur égale à $f(\xi_i)$, où ξ_i $(i = 1, 2, \ldots)$ est un point intermédiaire du i-ième

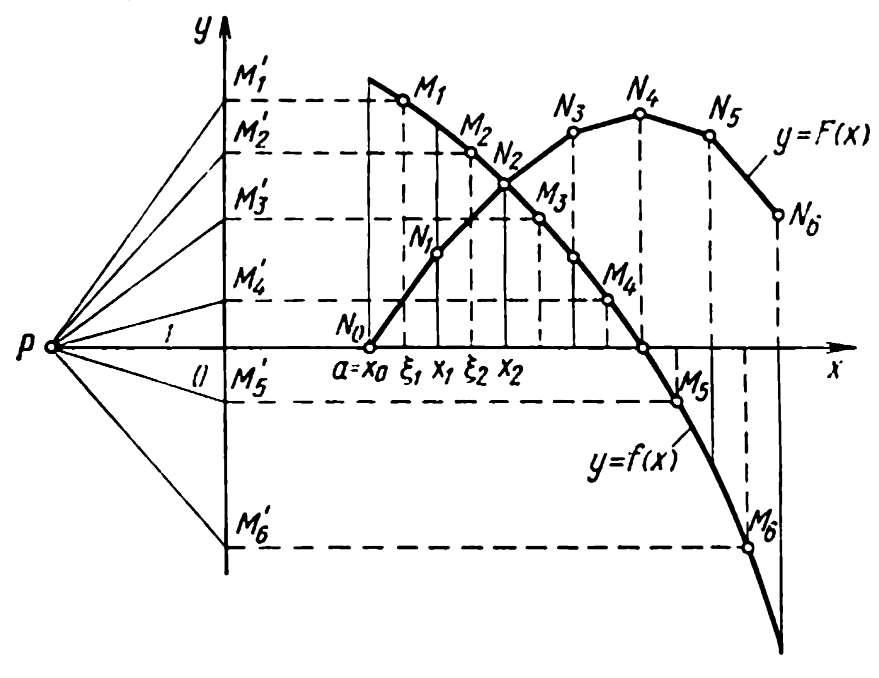


Fig. 77.

segment $[x_{i-1}, x_i]$. Posons donc

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = f(\xi_i)(x_i - x_{i-1}),$$

où

$$x_{i-1} \leqslant \xi_i \leqslant x_i \quad (i = 1, 2, \ldots).$$

Les valeurs de la primitive

$$F(x) = \int_{x_0}^{x} f(x)\,dx$$

aux points x_i peuvent être calculées par *additions successives*:

$F(x_0) = 0$;

$$F(x_i) = \int_{x_0}^{x_i} f(x)\,dx = \int_{x_0}^{x_{i-1}} f(x)\,dx + \int_{x_{i-1}}^{x_i} f(x)\,dx =$$

$$= F(x_{i-1}) + f(\xi_i)(x_i - x_{i-1}) \quad (i = 1, 2, \ldots). \qquad (1)$$

Soient $M_1(\xi_1, f(\xi_1))$, $M_2(\xi_2, f(\xi_2))$, ... les points respectifs de la courbe $y = f(x)$. En les projetant sur l'axe Oy on obtient les points M_1', M_2', ... (fig. 77).

Choisissons maintenant un pôle P tel que la distance $OP = 1$ et menons les rayons PM'_1, PM'_2, . . . La ligne cherchée $y = F(x)$ peut être remplacée approximativement par la ligne brisée $N_0N_1N_2N_3$ . . . aux sommets $N_0(x_0, 0)$, $N_1(x_1, F(x_1))$, $N_2(x_2, F(x_2))$, . . . Les éléments successifs de cette ligne brisée sont parallèles aux rayons correspondants, à savoir: $N_0N_1 \parallel PM'_1$; $N_1N_2 \parallel PM'_2$; $N_2N_3 \parallel PM'_3$; . . . En effet, en vertu de la formule (1), le coefficient angulaire de l'élément $N_{i-1}N_i$ est égal à

$$k = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}} = f(\xi_i),$$

mais par construction le coefficient angulaire du rayon OM'_i est

$$k'_i = \frac{f(\xi_i)}{1} = f(\xi_i).$$

Donc

$$N_{i-1}N_i \parallel OM'_i \qquad (i = 1, 2, \ldots).$$

Ainsi, la construction de la courbe d'une primitive $y = F(x)$ se ramène pratiquement à mener par le point $N_0(x_0, 0)$ la droite N_0N_1 parallèle au rayon OM'_1 jusqu'à l'intersection en N_1 avec la verticale $x = x_1$; par le point N_1 la droite N_1N_2 parallèle au rayon OM'_2 jusqu'à l'intersection au point N_2 avec la verticale $x = x_2$, etc.

Il convient d'indiquer qu'en appliquant la méthode d'intégration graphique considérée il n'est pas de rigueur de prendre les points x_i $(i = 0, 1, \ldots)$ équidistants. Pour améliorer la précision de la construction on recommande d'ajouter aux points x_i les points singuliers de la courbe de la fonction intégrée (zéros, points d'extrémum, points d'inflexion).

L'intégration graphique est dans le cas général peu précise. Aussi présente-t-elle de l'intérêt surtout lorsqu'il faut obtenir une idée générale sur l'intégrale de la fonction ou lorsque la fonction sous le signe somme est donnée graphiquement et nous ne connaissons pas son expression analytique.

§ 17*. Notion sur les formules de cubature

Les *formules de cubature* ou *formules des cubatures numériques* sont prévues pour le calcul numérique des intégrales doubles [1].

Soit la fonction $z = f(x, y)$ définie et continue dans un certain domaine borné σ (fig. 78). Dans ce domaine σ on choisit un système de points $M_i(x_i, y_i)$ $(i = 1, 2, \ldots, N)$. Pour calculer une intégrale double

$$\iint\limits_{(\sigma)} f(x, y)\, dx\, dy$$

on pose approximativement

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{N} A_i f(x_i, y_i). \qquad (1)$$

Pour trouver les coefficients A_i imposons la condition que la formule de cubature (1) soit exacte pour tout polynôme

$$P_n(x, y) = \sum_{k+l \leqslant n} c_{kl} x^k y^l, \qquad (2)$$

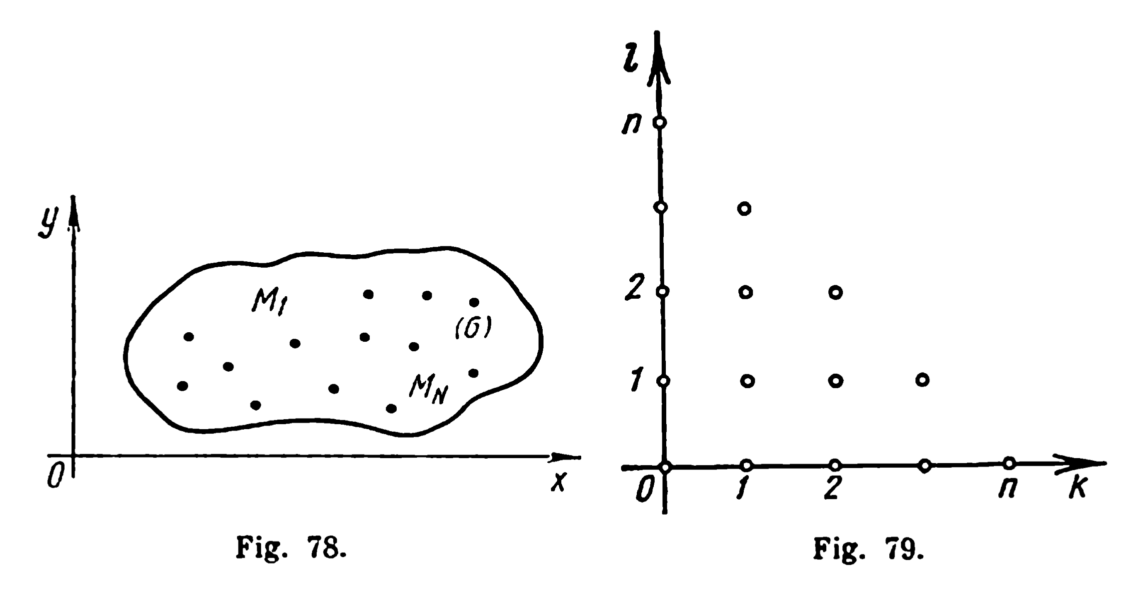


Fig. 78. Fig. 79.

dont le degré ne dépasse pas le nombre n donné. A cette fin il faut et il suffit que la formule (1) soit exacte pour les monômes

$$x^k y^l \quad (k,\ l = 0,\ 1,\ 2,\ \ldots,\ n;\ \ k + l \leqslant n).$$

Adoptant dans (1) $f(x, y) = x^k y^l$, on aura:

$$I_{kl} = \iint\limits_{(\sigma)} x^k y^l\,dx\,dy = \sum_{i=1}^{N} A_i x_i^k y_i^l \quad (k,\ l = 0,\ 1,\ 2,\ \ldots,\ n;\ k + l \leqslant n). \qquad (3)$$

Ainsi, dans le cas général, les coefficients A_i de la formule (1) peuvent être définis à partir d'un système d'équations linéaires (3).

Pour que le système (3) soit défini, il faut que le nombre d'inconnues N soit égal au nombre d'équations. Par suite, en composant le « réseau des exposants » (fig. 79) on obtient:

$$N = (n+1) + n + \ldots + 1 = \frac{(n+1)(n+2)}{2}.$$

Le problème délicat du choix le plus avantageux des points pour le domaine considéré reste sans être résolu.

On peut indiquer encore un procédé assez général de calcul d'une intégrale double. Supposons que le domaine d'intégration soit borné par des courbes représentatives de fonctions continues univoques

$$y = \varphi(x), \quad y = \psi(x)$$
$$(\varphi(x) \leqslant \psi(x))$$

et par deux verticales $x = a$, $x = b$ (fig. 80).

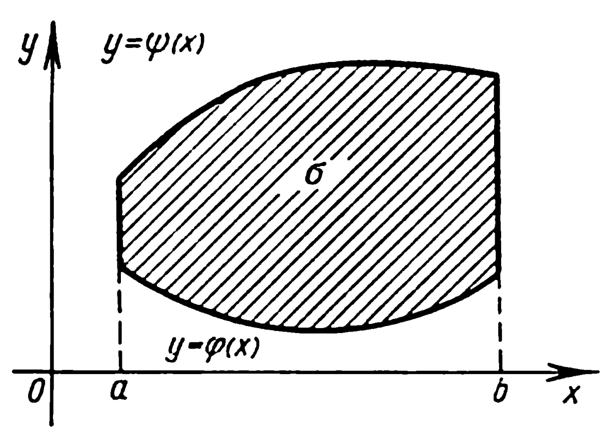


Fig. 80.

En plaçant dans l'intégrale double

$$I = \iint\limits_{(\sigma)} f(x, y)\,dx\,dy \qquad (4)$$

les limites d'intégration suivant des règles connues, on aura:

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \int\limits_a^b dx \int\limits_{\varphi(x)}^{\psi(x)} f(x, y)\,dy.$$

Soit

$$F(x) = \int\limits_{\varphi(x)}^{\psi(x)} f(x, y)\,dy. \qquad (5)$$

Alors

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \int\limits_a^b F(x)\,dx. \qquad (6)$$

Appliquant à l'intégrale simple du second membre de l'égalité (6) l'une des formules de quadrature on obtient:

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{n} C_i F(x_i), \qquad (7)$$

où $x_i \in [a, b]$ $(i = 1, 2, \ldots, n)$ et C_i sont des constantes. A leur tour les valeurs

$$F(x_i) = \int\limits_{\varphi(x_i)}^{\psi(x_i)} f(x_i, y)\,dy$$

peuvent s'obtenir également d'après certaines formules de quadrature

$$F(x_i) = \sum_{j=1}^{m_i} B_{ij} f(x_i, y_j),$$

où B_{ij} sont des constantes correspondantes.

On déduit de la formule (7):

$$\iint_{(\sigma)} f(x, y)\,dx\,dy = \sum_{i=1}^{n} \sum_{j=1}^{m_i} C_i B_{ij} f(x_i, y_j), \tag{8}$$

où C_i et B_{ij} sont des constantes connues.

L'interprétation géométrique de cette méthode est équivalente au calcul du volume I traduit par l'intégrale (4) à l'aide des sections transversales.

Avec des modifications correspondantes, les remarques générales relatives au calcul des intégrales simples (cf. § 10) gardent toute leur valeur pour les formules de cubature du type (8).

§ 18*. Formule de cubature de type Simpson

Supposons d'abord que le domaine d'intégration soit un rectangle

$$R\ \{a \leqslant x \leqslant A;\quad b \leqslant y \leqslant B\}$$

(fig. 81) dont les côtés sont parallèles aux axes de coordonnées. Divisons chacun des segments $[a, A]$ et $[b, B]$ en deux parties égales par les points

$$x_0 = a, \quad x_1 = a + h, \quad x_2 = a + 2h = A$$

et respectivement

$$y_0 = b, \quad y_1 = b + k, \quad y_2 = b + 2k = B,$$

où

$$h = \frac{A - a}{2}, \quad k = \frac{B - b}{2}.$$

On obtient ainsi au total neuf points (x_i, y_j) $(i, j = 0, 1, 2, \ldots$
$\ldots, 9)$. On a:

$$\iint_{(R)} f(x, y)\,dx\,dy = \int_a^A dx \int_b^B f(x, y)\,dy. \tag{1}$$

On en tire en calculant l'intégrale intérieure d'après la formule de quadrature de Simpson:

$$\iint f(x, y)\,dx\,dy = \int_a^A dx \cdot \frac{k}{3}[f(x, y_0) + 4f(x, y_1) + f(x, y_2)] =$$

$$= \frac{k}{3}\left[\int_a^A f(x, y_0)\,dx + 4\int_a^A f(x, y_1)\,dx + \int_a^A f(x, y_2)\,dx\right].$$

Appliquant encore une fois à chaque intégrale la formule de Simpson, on obtient:

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\{[f(x_0, y_0) + 4f(x_1, y_0) + f(x_2, y_0)] + \\ + 4[f(x_0, y_1) + 4f(x_1, y_1) + f(x_2, y_1)] + \\ + [f(x_0, y_2) + 4f(x_1, y_2) + f(x_2, y_2)]\}$$

ou

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \frac{hk}{9}\{[f(x_0, y_0) + f(x_2, y_0) + f(x_0, y_2) + \\ + f(x_2, y_2)] + 4[f(x_1, y_0) + f(x_0, y_1) + f(x_2, y_1) + \\ + f(x_1, y_2)] + 16f(x_1, y_1)\}. \quad (2)$$

Appelons la formule (2) *formule de cubature de Simpson*. Par conséquent,

$$\iint\limits_{(R)} f(x, y)\,dx\,dy = \\ = \frac{hk}{9}(\sigma_0 + 4\sigma_1 + 16\sigma_2), \quad (2')$$

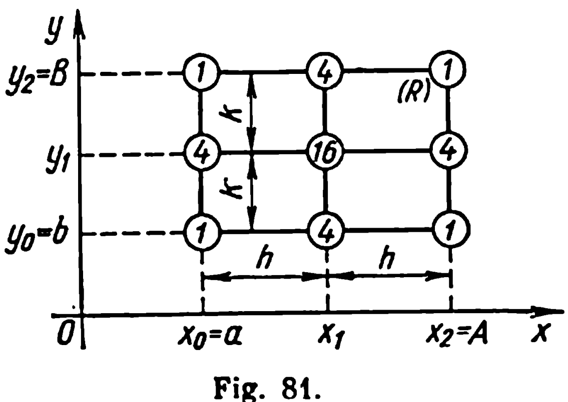


Fig. 81.

où σ_0 est la somme des valeurs de la fonction à intégrer $f(x, y)$ aux sommets du rectangle R; σ_1 la somme des valeurs de $f(x, y)$ au milieu des côtés du rectangle R; $\sigma_2 = f(x_1, y_1)$ la valeur de la fonction $f(x, y)$ au centre du rectangle R. Les multiplicités de ces valeurs sont représentées sur la figure 81.

Exemple 1. Appliquer la formule de cubature de Simpson pour calculer l'intégrale double [7]

$$I = \int\limits_4^{4,4}\int\limits_2^{2,6} \frac{dx\,dy}{xy}.$$

Solution. Prenons

$$h = \frac{4,4-4}{2} = 0,2 \quad \text{et} \quad k = \frac{2,6-2}{2} = 0,3.$$

Les valeurs correspondantes de la fonction sous le signe somme $z = \dfrac{1}{xy}$ sont portées sur le tableau 75.

Tableau 75

Calcul de l'intégrale double suivant la formule de Simpson

y_j \ x_i	4,0	4,2	4,4
2,0	0,125000	0,119048	0,113636
2,3	0,108696	0,103520	0,0988142
2,6	0,096154	0,0915751	0,0874126

L'application de la formule (2) donne

$$I=\frac{0,2\cdot 0,3}{9}[(0,125000+0,113636+0,096154+0,0874126)+$$
$$+4(0,119048+0,108696+0,0988142+0,0915751)+$$
$$+16\cdot 0,103520]=0,0250070.$$

La valeur exacte de cette intégrale double est:

$$\int_4^{4,4}\int_2^{2,6}\frac{dx\,dy}{xy}=\ln 1,3\cdot\ln 1,1=0,0953108\cdot 0,262364=0,0250061.$$

Donc l'erreur de troncature

$$\Delta=|0,0250061-0,0250070|=0,0000009\approx 10^{-6}.$$

Si les dimensions du rectangle R $\{a\leqslant x\leqslant A;\ b\leqslant y\leqslant B\}$ sont grandes, pour améliorer la précision de la formule (2) on divise

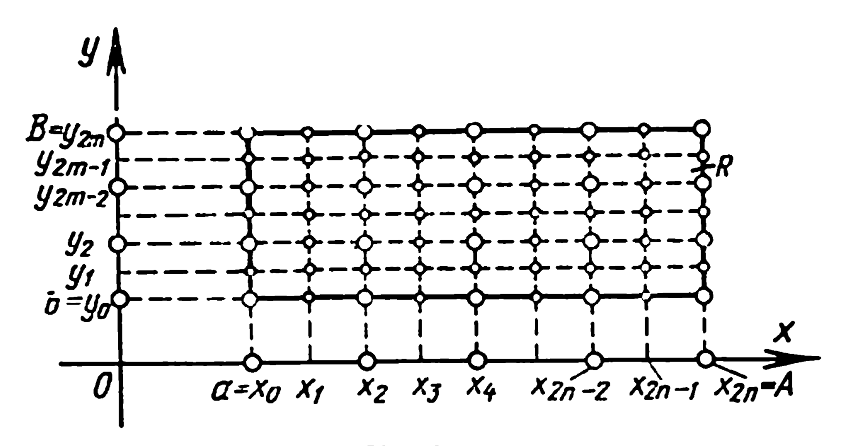


Fig. 82.

le domaine R en un système de rectangles pour appliquer à chacun de ces rectangles la formule de Simpson.

Supposons que nous ayons divisé les côtés du rectangle R respectivement en n et m parties égales; il en résulte un réseau relati-

vement lâche de nm rectangles (sur la fig. 82 les sommets de ces rectangles sont marqués par des ronds plus grands). Divisons à son tour chacun de ces rectangles en quatre parties égales. Adoptons que les smomets de ce réseau serré des rectangles sont des points M_{ij} de la formule de cubature.

Soit

$$h=\frac{A-a}{2n}$$

et

$$k=\frac{B-b}{2m}.$$

Les coordonnées des points du réseau sont alors les suivantes:

$$x_i=x_0+ih \quad (x_0=a;\; i=0,\ 1,\ 2,\ \ldots,\ 2n)$$

et

$$y_j=y_0+jk \quad (y_0=b;\; j=0,\ 1,\ 2,\ \ldots,\ 2m).$$

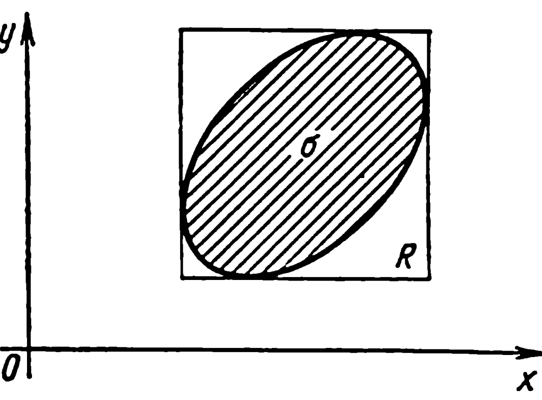


Fig. 83.

Pour abréger, introduisons la notation

$$f(x_i,\ y_i)=f_{ij}.$$

En appliquant la formule (2) à chacun des rectangles du réseau lâche, on aura (fig. 82):

$$\iint\limits_{(R)} f(x,\ y)\,dx\,dy=\frac{hk}{9}\sum_{i=0}^{n}\sum_{j=0}^{m}\left[(f_{2i,\,2j}+f_{2i+2,\,2j}+f_{2i+2,\,2j+2}+f_{2i,\,2j+2})+\right.$$
$$\left.+4\,(f_{2i+1,\,2j}+f_{2i+2,\,2j+1}+f_{2i+1,\,2j+2}+f_{2i,\,2j+1})+16f_{2i+1,\,2j+1}\right].$$

On en tire finalement après réduction des termes semblables:

$$\iint\limits_{(R)} f(x,\ y)\,dx\,dy=\frac{hk}{9}\sum_{i=0}^{2n}\sum_{j=0}^{2m}\lambda_{ij}f_{ij}, \tag{3}$$

où les coefficients λ_{ij} sont les éléments correspondants de la matrice

$$\Lambda=\begin{bmatrix} 1 & 4 & 2 & 4 & 2 & \ldots & 4 & 2 & 4 & 1\\ 4 & 16 & 8 & 16 & 8 & \ldots & 16 & 8 & 16 & 4\\ 2 & 8 & 4 & 8 & 4 & \ldots & 8 & 4 & 8 & 2\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 2 & 8 & 4 & 8 & 4 & \ldots & 8 & 4 & 8 & 2\\ 4 & 16 & 8 & 16 & 8 & \ldots & 16 & 8 & 16 & 4\\ 1 & 4 & 2 & 4 & 2 & \ldots & 4 & 2 & 4 & 1 \end{bmatrix}.$$

Si le domaine d'intégration σ est curviligne, on construit le rectangle $R \supset \sigma$, dont les côtés sont parallèles aux axes des coordonnées (fig. 83). Considérons la fonction auxiliaire

$$f^*(x, y) = \begin{cases} f(x, y), & \text{si } (x, y) \in \sigma; \\ 0, & \text{si } (x, y) \in R - \sigma. \end{cases}$$

Il est clair que dans ce cas on a:

$$\iint\limits_{(\sigma)} f(x, y)\,dx\,dy = \iint\limits_{(R)} f^*(x, y)\,dx\,dy.$$

Le calcul approché de cette dernière intégrale peut se faire d'après la formule de cubature générale (3).

BIBLIOGRAPHIE

1. *Ch. Mikéladzé.* Méthodes numériques de l'analyse mathématique. Gostekhizdat, 1953, chapitres XIII, XVIII.
2. *W. E. Milne.* Numerical calculus. Princeton University Press, Princeton, 1949, chapitre IV.
3. *S. Nikolski.* Formules de quadrature. Fizmatguiz, Moscou, 1958.
4. *A. Markov.* Calcul des différences finies, 2[e] éd. Matézis, 1911, chapitre V.
5. *J. F. Steffensen.* Iinterpolation. Baltimore. 1927.
6. *I. Bérézine, N. Jidkov.* Méthodes de calcul. Fizmatguiz, Moscou, 1959, tome I, chapitre III.
7. *J. B. Scarborough.* Numerical Mathematical Analysis. John Hopkins, 1950 (2[e] éd.).
8. *A. Krylov.* Conférences sur le calcul approché, 2[e] éd. Nouvelles de l'Académie des Sciences de l'U.R.S.S., 1933, chapitre III.
9. *V. Krylov.* Calcul approché des intégrales. Fizmatguiz, Moscou, 1959.
10. *M. G. Salvadori.* Numerical methods in engineering. New York, Prentice-Hall, 1952.
11. *G. Fichtengoltz.* Cours de calcul différentiel et intégral, 1948, Gostekhizdat, t. 2, chapitres IX, XIII.

CHAPITRE XVII

MÉTHODE DE MONTE-CARLO

§ 1. Principe de la méthode

Le mode usuel de résolution d'un problème consiste à indiquer un algorithme (la succession des opérations) qui permet de trouver la valeur f exacte ou avec une précision donnée. Notamment, si l'on désigne par $f_1, f_2, \ldots, f_n, \ldots$ les résultats correspondants des opérations successives, alors

$$f = \lim_{n \to \infty} f_n. \tag{1}$$

et dans le cas d'un nombre fini d'opérations le processus s'arrête à un certain pas. Le processus de calcul est dans ce cas strictement déterministe: en l'absence d'erreurs, deux calculateurs différents aboutissent au même résultat.

Toutefois, il existe des problèmes dans lesquels la construction des algorithmes de ce type est pratiquement impossible ou l'algorithme lui-même s'avère trop compliqué. On recourt alors souvent à la simulation du principe mathématique ou physique du problème et on applique les lois des grands nombres de la théorie des probabilités. Les estimations $f_1, f_2, \ldots, f_n, \ldots$ de la grandeur cherchée f s'obtiennent par traitement statistique des données fournies par les résultats de certaines *expériences aléatoires* multiples. Dans ces conditions il faut que la variable aléatoire f_n converge en probabilité pour $n \to \infty$ vers la grandeur cherchée f [1], [2], c'est-à-dire que pour tout $\varepsilon > 0$ on ait la relation limite

$$\lim_{n \to \infty} P(|f - f_n| < \varepsilon) = 1, \tag{2}$$

où P désigne la probabilité correspondante.

Le choix de la grandeur f_n est conditionné par des particularités concrètes du problème. Par exemple, on entend souvent par grandeur cherchée f la probabilité d'un certain événement aléatoire (ou, pour plus de généralité, l'espérance mathématique d'une certaine variable aléatoire). Alors, la fréquence f_n d'un événement dans n expériences aléatoires (ou, respectivement, la moyenne empirique des valeurs d'une variable aléatoire) peut être considérée sous des

hypothèses très lâches comme une estimation probabiliste de la variable cherchée. D'autres variantes sont également possibles. Constatons que dans ce cas le processus de calcul est non déterministe, puisqu'il est défini par les résultats des expériences aléatoires.

Les modes de résolution des problèmes faisant appel aux variables aléatoires ont reçu le nom général de la *méthode de Monte-Carlo.* Plus précisément par méthode de Monte-Carlo [3], [4], [5], [6] on entend l'ensemble des procédés qui permettent d'obtenir la solution des problèmes mathématiques et physiques à l'aide des expériences aléatoires multiples. Les estimations de la grandeur cherchée se déduisent statistiquement et ont un caractère probabiliste. Dans la pratique les expériences aléatoires sont remplacées par certains calculs appliqués aux *nombres aléatoires* (cf. § 2).

L'utilisation efficace de la méthode de Monte-Carlo est devenue possible grâce aux calculateurs électroniques rapides, car pour obtenir des estimations suffisamment exactes de la grandeur cherchée, il faut réaliser le calcul d'un très grand nombre de cas particuliers et dépouiller ensuite la statistique d'un volume énorme de données numériques. Remarquons qu'en utilisant la méthode de Monte-Carlo aucun besoin n'est de connaître les relations précises des grandeurs données et recherchées du problème, il suffit de dégager seulement l'ensemble des conditions qui définissent la manifestation du phénomène observé. Cette circonstance rend possible l'application de la méthode de Monte-Carlo aux problèmes logiques.

Voici les problèmes mathématiques pour lesquels on a mis au point la méthode de Monte-Carlo: résolution des systèmes d'équations linéaires; inversion des matrices; recherche des valeurs propres et des vecteurs propres d'une matrice; calcul des intégrales multiples; résolution du problème de Dirichlet; résolution des équations fonctionnelles de divers types, etc. La méthode de Monte-Carlo permet également de résoudre des problèmes de physique nucléaire. Notons que pour un même problème concret, le schéma de l'application de la méthode peut être nettement différent.

Dans ce chapitre nous allons étudier le calcul des intégrales multiples et la résolution des systèmes d'équations linéaires par la méthode de Monte-Carlo. Pour se renseigner sur les autres problèmes indiqués il faut se référer aux ouvrages appropriés (cf. par exemple, [3], bibliographie, ainsi que [6]).

§ 2. Nombres aléatoires

Dans la pratique de la méthode de Monte-Carlo les expériences aléatoires sont remplacées généralement par un échantillonnage des *nombres aléatoires.*

D é f i n i t i o n 1. Une grandeur ou une variable est dite *aléatoire* si sa valeur dépend d'un événement aléatoire.

La variable aléatoire X est définie par la loi de répartition

$$P(X < x) = \Phi(x),$$

où x est un nombre réel quelconque et $\Phi(x)$ une fonction connue (*fonction de répartition*). Les valeurs de la variable aléatoire s'appellent *nombres aléatoires.*

D é f i n i t i o n 2. Si une variable aléatoire est munie d'une loi de répartition donnée [1], [2] (uniforme, normale, etc.), on dit que les nombres aléatoires correspondants sont *répartis d'après cette loi.*

Soient les nombres $x_1, x_2, \ldots, x_n, \ldots$ les valeurs d'une même variable aléatoire X fournies par des épreuves indépendantes à conditions répétées. Alors, la *suite des nombres aléatoires*

$$\{x_n\} \tag{1}$$

est dite *aléatoire*, à loi de répartition correspondante. Dans ce qui suit nous allons étudier en règle générale des suites aléatoires (1) *à répartition uniforme* sur un segment unité $0 \leqslant x \leqslant 1$. Si (a, b) est un intervalle quelconque * du segment $[0, 1]$ et $\nu_n = \nu_n(a, b)$, le nombre d'éléments de la sous-suite finie $x_1, x_2, \ldots, x_n$ appartenant à l'intervalle (a, b), alors pour la suite (1) à répartition uniforme on a la relation limite suivante

$$\lim_{n \to \infty} \frac{\nu_n(a, b)}{n} = b - a, \tag{2}$$

c'est-à-dire la *fréquence relative limite de la suite* $\{x_n\}$ *à répartition uniforme sur* $[0, 1]$ *pour tout intervalle partiel* (a, b) *est égale à la longueur de cet intervalle avec la probabilité* 1.

Si la suite aléatoire $\{x_n\}$ est répartie uniformément sur le segment $[0, 1]$, la transformation linéaire

$$y_n = A + (B - A)\, x_n \quad (n = 1, 2, \ldots), \tag{3}$$

où A et B sont des nombres donnés, conduit à la suite aléatoire $\{y_n\}$ répartie uniformément sur le segment $[A, B]$.

Dans le cas général, une suite aléatoire $\{x_n\}$ répartie uniformément sur le segment $[0, 1]$ permet de construire une suite aléatoire $\{y_n\}$ à loi de répartition donnée $\Phi(y)$.

Soit

$$\Phi(y) = \int_{-\infty}^{y} \varphi(t)\, dt$$

* Les extrémités a et b peuvent par convention être ou ne pas être incluses dans l'intervalle (a, b).

la fonction de répartition correspondante *, où $\varphi(t)$ est la *densité de probabilité*.

Pour simplifier supposons que la fonction

$$x = \Phi(y)$$

soit continue et strictement monotone (fig. 84). Alors, en définissant y_n d'après l'équation

$$x_n = \Phi(y_n) \quad (n = 1, 2, \ldots),$$

on obtient pour tout x_n la suite aléatoire $\{y_n\}$ munie de la loi de répartition donnée $\Phi(y)$. Par construction, la suite $\{y_n\}$ vérifie

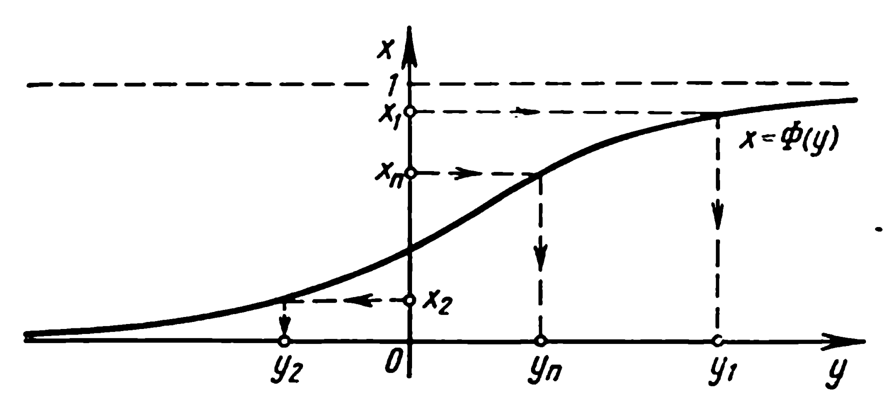


Fig. 84.

avec la probabilité 1 la relation limite

$$\lim_{n\to\infty} \frac{\tilde{\nu}_n(a, b)}{n} = \int_a^b \varphi(y)\, dy, \tag{4}$$

où $\tilde{\nu}_n(a, b)$ est le nombre d'éléments d'une sous-suite finie $y_1, \ldots$ $\ldots, y_n$, appartenant à l'intervalle arbitraire (a, b).

En particulier, en posant

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}},$$

on obtient de cette façon la suite aléatoire canonique $\{y_n\}$ obéissant à la loi normale (gaussienne) et associée à la variable aléatoire Y d'espérance mathématique $MY = 0$ et de variance $DY = 1$. La transformation linéaire

$$z_n = \sigma y_n + c \quad (n = 1, 2, \ldots)$$

donne une suite aléatoire $\{z_n\}$ à répartition normale qui correspond à la variable aléatoire Z telle que l'espérance mathématique $MZ = c$ et la variance $DZ = \sigma^2$.

* Si y_n $(n = 1, 2, \ldots)$ sont contenues dans le segment limité $A \leqslant y \leqslant B$, alors on pose généralement $\varphi(y) = 0$ avec $y \,\overline{\in}\, [A, B]$.

§ 3. Méthodes d'obtention des nombres aléatoires

Pour élaborer des nombres aléatoires on peut utiliser les résultats de processus physiques aléatoires (par exemple, le jet des dés, la rotation de la roulette, le scintillement du compteur Geiger, le bruit des transmissions électriques, etc.). Il existe également des tables toutes prêtes des nombres aléatoires (cf., par exemple, [7], [8]).

En toute rigueur, utilisant des dispositifs mécaniques, pour obtenir des nombres aléatoires, on ne peut pas être tout à fait sûr que les événements aléatoires considérés ont une répartition de probabilité donnée. C'est pourquoi on soumet généralement les données obtenues à une « vérification statistique par le hasard ». Dans ce sens il est plus sûr d'employer des nombres aléatoires tabulés pour lesquels cette vérification est déjà faite; pourtant les nombres aléatoires tabulés présentent de grands inconvénients pour le traitement des problèmes sur des machines digitales [9].

Pour résoudre les problèmes par la méthode de Monte-Carlo il faut avoir à sa disposition une grande quantité de nombres aléatoires. Dans la pratique le plus commode est d'obtenir ces nombres avec des *détecteurs spéciaux* couplés à une machine. Leur fonctionnement est réglé par des processus physiques aléatoires (par exemple, par désintégration radioactive, bruits des tubes électroniques, etc.) [9].

La reproduction des nombres aléatoires associés au modèle théorique donné étant un processus délicat et compliqué, on se borne souvent en pratique à l'obtention de ce qu'on appelle les nombres pseudo-aléatoires qui *grosso modo* ressemblent aux nombres aléatoires correspondants. Les nombres pseudo-aléatoires sont tirés à partir des algorithmes assez complexes. Dans ce qui suit, par « nombre aléatoire » nous allons entendre les nombres de ces deux types s'ils ne présentent pas de différence substantielle.

Indiquons certains procédés bien simples pour obtenir des nombres aléatoires, au sens généralisé, uniformément répartis sur le segment [0, 1]. Supposons pour simplifier que ces nombres sont des fractions décimales propres du nombre fixe, s par exemple, de décimales significatives (fraction décimale à *s rangs*), c'est-à-dire pouvant être mise sous la forme

$$x = \frac{\alpha_1}{10} + \frac{\alpha_2}{10^2} + \ldots + \frac{\alpha_s}{10^s}, \qquad (1)$$

où α_i ($i = 1, 2, \ldots, s$) sont les chiffres de ces nombres, prenant les valeurs 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

Pour former le tableau des nombres aléatoires de la forme (1), uniformément répartis sur le segment [0, 1], il suffit d'indiquer les modes d'obtention des chiffres α_i en respectant les conditions suivantes:

a) α_i est un *échantillon aléatoire* du système des nombres 0 à 9, toutes les valeurs indiquées étant équiprobables et indépendantes;

b) le choix des chiffres précédents $\alpha_1, \ldots, \alpha_i$ n'influe nullement sur celui du chiffre suivant α_{i+1}.

Pour obtenir un nombre aléatoire à s rangs, cet échantillonnage est repris s fois.

Il existe plusieurs procédés pour réaliser le système de sélection vérifiant les conditions a) et b). Examinons certains d'entre eux.

1. Plaçons dans une urne dix boules identiques numérotées de 0 à 9. Tirons successivement de l'urne une boule et inscrivons son numéro α. Après chaque tirage la boule est remise dans l'urne et, avant chaque tirage consécutif, toutes les boules dans l'urne sont brassées.

2. On jette simultanément deux dés. Si n_1 et n_2 sont les chiffres amenés ($n_1, n_2 = 1, 2, 3, 4, 5, 6$) respectivement par le premier et le deuxième dé (les deux dés doivent être différents), le chiffre successif α du nombre aléatoire est pris égal au reste de la division de la somme $6(n_1 - 1) + n_2$ par 10, où $n_1 < 6$, c'est-à-dire α est un entier non négatif inférieur à 10 qui vérifie la congruence *

$$6(n_1 - 1) + n_2 \equiv \alpha \pmod{10}. \tag{2}$$

Si $n_1 = 6$, on jette encore une fois les dés. La formule (2) entraîne que le chiffre α peut à probabilité égale prendre une valeur quelconque de 0 à 9 (cf. [7]).

3. On prend un entier à s chiffres. Ce nombre est élevé au carré, puis on choisit dans le nombre obtenu s chiffres moyens; ensuite le processus est repris. Si s est suffisamment grand, par exemple $s \geqslant 10$, les chiffres choisis peuvent être pris à chaque étape comme décimales des nombres pseudo-aléatoires à s rangs [3].

Pour obtenir une suite de nombres pseudo-aléatoires on peut également multiplier un nombre de plusieurs chiffres par un même nombre et en tirer les chiffres moyens ou élever au carré un nombre de plusieurs chiffres et calculer le reste de la division du résultat par un nombre premier suffisamment grand.

4. Une suite pseudo-aléatoire $\{x_n\}$ s'obtient à l'aide du processus [10]

$$x_n = 2^{-42} u_n,$$

où

$$u_0 = 1, \quad u_{n+1} \equiv 5^{17} u_n \pmod{2^{42}}.$$

5. On utilise le développement décimal d'un nombre irrationnel positif

$$\omega = \beta_0, \beta_1, \beta_2 \ldots \beta_s \ldots = \beta_0 + (\omega),$$

---

* L'écriture $a \equiv b \pmod{k}$ (a, b, k, entiers) signifie que la différence $a - b$ est divisible par k.

où β_0 est la partie entière du nombre ω et (ω) sa partie fractionnaire.

Pour obtenir une suite aléatoire $\{x_n\}$, on pose:

$$x_n = (n\omega) \quad (n = 1, 2, \ldots).$$

S'il faut obtenir une suite aléatoire composée de nombres à s rangs, dans les nombres $(n\omega)$ on se borne aux rangs correspondants.

Pour résoudre certains problèmes il faut avoir à sa disposition plusieurs suites aléatoires

$$\{x_n^{(1)}\}, \{x_n^{(2)}\}, \ldots, \{x_n^{(m)}\}.$$

Dans ce cas on choisit m nombres irrationnels positifs $\omega_1, \omega_2, \ldots$ $\ldots, \omega_m$ linéairement indépendants sur le corps des rationnels pour admettre

$$x_n^{(k)} = (n\omega_k) \quad (k = 1, 2, \ldots, m; \ n = 1, 2, \ldots).$$

On peut également prendre une suite aléatoire uniformément répartie $\{x_n\}$ et en tirer m échantillons:

$$\{x_1, x_{m+1}, x_{2m+1}, \ldots\},$$
$$\{x_2, x_{m+2}, \ldots x_{2m+2}, \ldots\},$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\{x_m, x_{2m}, x_{3m}, \ldots\},$$

en prenant les nombres non pas l'un après l'autre, mais de m à m. Il est clair que de cette façon on aura m sous-suites réparties uniformément.

Ces méthodes ainsi que bien d'autres ont servi pour dresser des tables des nombres aléatoires. Dans le cas général on donne dans ces tables les décimales aléatoires; on s'en sert pour construire des nombres aléatoires ayant un nombre déterminé de décimales. A titre d'exemple voici une partie d'une telle table (cf. [7]) à cinq décimales (tableau 76).

Tableau 76

Nombres aléatoires répartis uniformément sur le segment [0, 1]

0,57705	0,35483	0,11578	0,65339	0,66674
0,71618	0,09393	0,93045	0,93382	0,99279
0,73710	0,30304	0,93011	0,05758	0,24202
0,70131	0,55186	0,42844	0,00336	0,94010
0,16961	0,64003	0,52906	0,88222	0,60981
0,53324	0,20514	0,09461	0,98585	0,13094
0,43166	0,00188	0,99602	0,52103	0,35193
0,26275	0,55709	0,69962	0,91827	0,64560
0,05926	0,86977	0,31311	0,07069	0,64559
0,66289	0,31303	0,27004	0,13928	0,68008

§ 4. Calcul des intégrales multiples par la méthode de Monte-Carlo

Soit la fonction

$$y = f(x_1, x_2, \ldots, x_m)$$

continue dans un domaine fermé S; calculer l'intégrale m-uple

$$I = \underset{(S)}{\iint \ldots \int} f(x_1, x_2, \ldots, x_m)\, dx_1\, dx_2 \ldots dx_m. \qquad (1)$$

Géométriquement le nombre I est le volume de dimension $(m+1)$ d'un cylindroïde * droit dans l'espace $Ox_1x_2 \ldots x_my$ de base S et borné supérieurement par la surface donnée $y = f(x)$, où $x = (x_1, x_2, \ldots, x_m)$ (fig. 85).

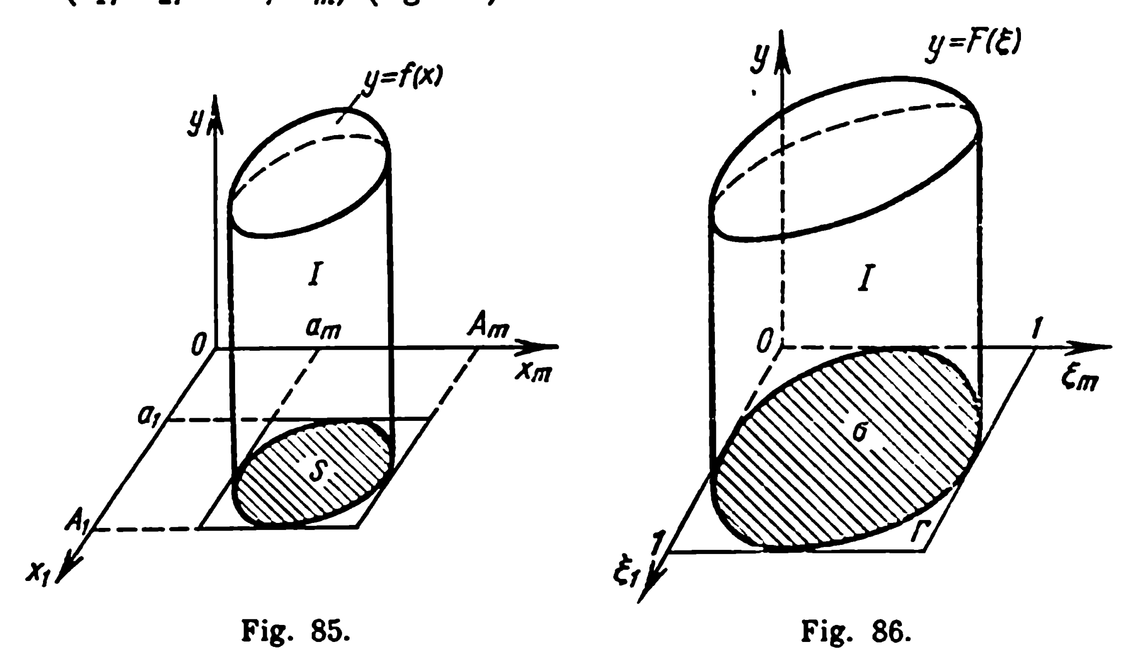


Fig. 85. Fig. 86.

Transformons l'intégrale (1) de façon que le nouveau domaine d'intégration soit intérieur à un cube unité de dimension m. Soit le domaine S intérieur à un parallélépipède de dimension m

$$a_i \leqslant x_i \leqslant A_i \quad (i = 1, 2, \ldots, m). \qquad (2)$$

Faisons le changement de variables

$$x_i = a_i + (A_i - a_i)\,\xi_i \quad (i = 1, 2, \ldots, m). \qquad (3)$$

Il est alors évident que le parallélépipède de dimension m (2) se transforme en un cube unité de dimension m

$$0 \leqslant \xi_i \leqslant 1 \quad (i = 1, 2, \ldots, m) \qquad (4)$$

* Plus précisément, le volume algébrique où l'on suppose que les parties du cylindroïde au-dessus de l'hyperplan $Ox_1x_2, \ldots, x_m$ aient une mesure positive et au-dessous, une mesure négative.

et, par conséquent, le nouveau domaine d'intégration σ, qui s'obtient suivant les règles usuelles est intérieur à ce cube (fig. 86).

En calculant le jacobien de la transformation, on aura:

$$\frac{D(x_1, x_2, \ldots, x_m)}{D(\xi_1, \xi_2, \ldots, \xi_m)} = \begin{vmatrix} A_1 - a_1 & 0 & \ldots & 0 \\ 0 & A_2 - a_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & A_m - a_m \end{vmatrix} = $$

$$= (A_1 - a_1)(A_2 - a_2) \ldots (A_m - a_m).$$

Ainsi

$$I = \iint\limits_{(\sigma)} \ldots \int F(\xi_1, \xi_2, \ldots, \xi_m)\, d\xi_1\, d\xi_2 \ldots d\xi_m, \tag{5}$$

où

$$F(\xi_1, \xi_2, \ldots, \xi_m) = (A_1 - a_1)(A_2 - a_2) \ldots (A_m - a_m) f(a_1 + \\ + (A_1 - a_1)\xi_1, a_2 + (A_2 - a_2)\xi_2, \ldots, a_m + (A_m - a_m)\xi_m).$$

En introduisant les notations

$$\xi = (\xi_1, \xi_2, \ldots, \xi_m)$$

et

$$d\sigma = d\xi_1\, d\xi_2 \ldots d\xi_m,$$

écrivons l'intégrale (5) sous une forme abrégée:

$$I = \iint\limits_{(\sigma)} \ldots \int F(\xi)\, d\sigma. \tag{5'}$$

Nous indiquerons deux méthodes de calcul de l'intégrale (5') par la méthode des expériences aléatoires.

P r e m i è r e m é t h o d e. Choisissons m suites aléatoires indépendantes uniformément réparties sur le segment [0, 1]:

$$\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_n^{(1)}, \ldots;$$
$$\xi_1^{(2)}, \xi_2^{(2)}, \ldots, \xi_n^{(2)}, \ldots;$$
$$\ldots \ldots \ldots \ldots$$
$$\xi_1^{(m)}, \xi_2^{(m)}, \ldots, \xi_n^{(m)}, \ldots$$

Les points $M_i(\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(m)})$ ($i = 1, 2, \ldots$) peuvent être considérés comme aléatoires. En choisissant un nombre N suffisamment grand de points $M_1, M_2, \ldots, M_N$, vérifions quels sont ceux qui appartiennent au domaine σ (*première espèce*) et ceux qui n'y appartiennent pas (*deuxième espèce*). Soit (fig. 87)

$$1) \quad M_i \in \sigma \quad \text{pour} \quad i = 1, 2, \ldots, n \tag{6}$$

et

$$2)\quad M_i \,\bar{\in}\, \sigma \quad \text{pour} \quad i = n+1,\ n+2,\ \ldots,\ N \qquad (6')$$

(par commodité nous changerons ici la numérotation des points). Remarquons qu'il faut convenir à l'avance si les points de la frontière Γ ou certains d'entre eux appartiennent au domaine σ ou non. Dans le cas général, lorsque la frontière Γ est lisse, cela n'a pas d'importance; mais dans des cas particuliers cette question doit être tranchée en tenant compte des conditions concrètes.

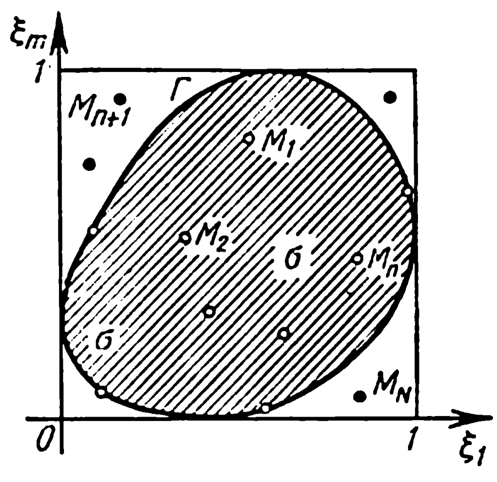


Fig. 87.

En prenant un nombre n suffisamment grand de points $M_i \in \sigma$, on peut poser approximativement

$$y_{\text{moy}} = \frac{1}{n} \sum_{i=1}^{n} F(M_i),$$

d'où il résulte la formule de l'intégrale cherchée

$$I = y_{\text{moy}}\sigma = \frac{\sigma}{n} \sum_{i=1}^{n} F(M_i), \qquad (7)$$

où par σ on entend le volume de dimension m du domaine d'intégration σ. Si le calcul du volume σ est difficile, on peut poser

$$\sigma \approx \frac{n}{N},$$

d'où

$$I \approx \frac{1}{N} \sum_{i=1}^{n} F(M_i).$$

Dans le cas particulier, lorsque σ est un cube unité ($\sigma = 1$), la vérification devient superflue, c'est-à-dire $n = N$ et on a simplement:

$$I = \frac{1}{N} \sum_{i=1}^{N} F(M_i).$$

Pour vérifier les conditions (6) et (6′) on part dans les cas courants de la donnée analytique de la frontière Γ du domaine σ. Dans le cas le plus simple, lorsque la surface Γ est donnée par l'équation

$$\varphi(\xi) = 0, \qquad (8)$$

où avec $\varphi(\xi) < 0$ le point $\xi \in \sigma$ et avec $\varphi(\xi) > 0$ le point $\xi \,\bar{\in}\, \sigma$, on a: 1) si $\varphi(M_i) < 0$, le point M_i est de première espèce et 2) si

$\varphi(M_i) > 0$, le point M_i est de deuxième espèce. Les points M_i tels que $\varphi(M_i) = 0$ sont attribués à la première ou à la deuxième espèce par convention. Constatons que l'équation (8) peut être remplacée par n'importe quelle équation équivalente, ce qui rend parfois les calculs bien plus faciles. Ainsi, pour un cercle il est commode de remplacer l'inégalité

$$x^2+y^2-x-y+\frac{1}{4}\leqslant 0$$

par une inégalité équivalente

$$\left(x-\frac{1}{2}\right)^2+\left(y-\frac{1}{2}\right)^2\leqslant\frac{1}{4},$$

la deuxième inégalité étant plus simple à vérifier.

Si le domaine σ est donné par les inégalités

$$\left.\begin{aligned} &\underline{\xi}_1\leqslant\xi_1\leqslant\bar{\xi}_1,\\ &\underline{\xi}_2(\xi_1)\leqslant\xi_2\leqslant\bar{\xi}_2(\xi_1),\\ &\dots\dots\dots\dots\dots\dots\dots\dots\\ &\underline{\xi}_m(\xi_1,\ \dots,\ \xi_{m-1})\leqslant\xi_m\leqslant\bar{\xi}_m(\xi_1,\ \dots,\ \xi_{m-1}), \end{aligned}\right\}\qquad(9)$$

l'appartenance d'un point aléatoire $M(\xi_1, \xi_2, \dots, \xi_m)$ à la première ou à la deuxième espèce s'établit en vérifiant si ces inégalités sont respectées.

Tableau 77

Schéma déterminant si le point aléatoire $M(\xi_1, \dots, \xi_m)$ appartient au domaine (9)

ξ_1	$\underline{\xi}_1$	$\bar{\xi}_1$	ε_1	ξ_2	$\underline{\xi}_2$	$\bar{\xi}_2$	ε_2

...	ξ_m	$\underline{\xi}_m$	$\bar{\xi}_m$	ε_m	ε	ν

En pratique, le plus commode est de recourir au schéma du tableau 77. Ici

$$\varepsilon_i = \begin{cases} 1, & \text{si } \xi_i \in [\underline{\xi}_i, \overline{\xi}_i] \\ 0, & \text{si } \xi_i \overline{\in} [\underline{\xi}_i, \overline{\xi}_i], \end{cases}$$

$(i = 1, 2, \ldots, m)$ et $\varepsilon = \varepsilon_1\varepsilon_2 \ldots \varepsilon_m$. Il est évident que

$$\text{si } \varepsilon = 1, \text{ alors } M \in \sigma;$$
$$\text{si } \varepsilon = 0, \text{ alors } M \overline{\in} \sigma.$$

Remarquons que si $\varepsilon_j = 0$ $(j < m)$ aucun besoin n'est de calculer les valeurs ultérieures $\varepsilon_{j+1}, \ldots, \varepsilon_m$, puisqu'elles n'influent pas sur le résultat définitif. La valeur de la fonction $y = F(M)$ ne se calcule que pour les points M tels que $\varepsilon = 1$. Ensuite, pour calculer l'intégrale I, on utilise la formule (7).

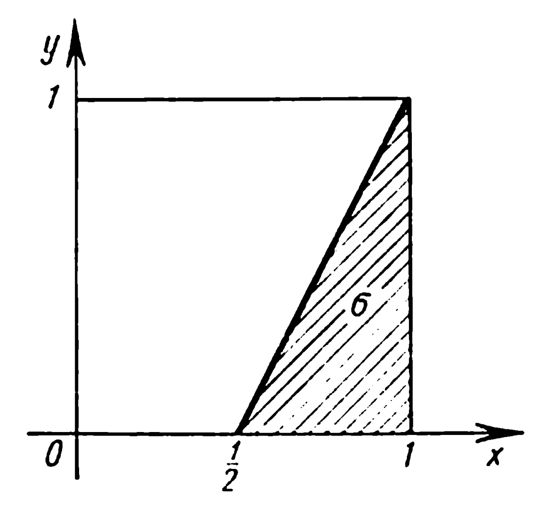


Fig. 88.

E x e m p l e. Calculer approximativement par la méthode de Monte-Carlo l'intégrale

$$I = \iint\limits_{(\sigma)} (x^2 + y^2)\, dx\, dy, \qquad (10)$$

où le domaine d'intégration σ est défini par les inégalités suivantes:

$$\left.\begin{array}{l} \frac{1}{2} \leqslant x \leqslant 1, \\ 0 \leqslant y \leqslant 2x - 1 \end{array}\right\} \qquad (\sigma)$$

(fig. 88).

S o l u t i o n. L'intégrale (10) est donnée sous une forme réduite, c'est-à-dire le domaine d'intégration σ est intérieur au carré unité

$$0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1.$$

Pour résoudre le problème faisons appel au tableau 76 des nombres aléatoires en considérant chaque couple successif des nombres du tableau comme les coordonnées correspondantes x et y du point aléatoire $M(x, y)$. Le calcul ayant un caractère illustratif, bornons-nous à $N = 20$ points aléatoires en arrondissant pour simplifier leurs coordonnées à trois chiffres décimaux. Les résultats du calcul sont portés sur le tableau 78, où nous avons posé

$$\underline{x} = \frac{1}{2}, \ \overline{x} = 1;$$
$$\underline{y}(x) = 0, \ \overline{y}(x) = 2x - 1;$$
$$z = x^2 + y^2.$$

Tableau 78

Calcul de l'intégrale double (10) par la méthode de Monte-Carlo

x	$\underline{x}$	$\bar{x}$	ε_1	y	$\underline{y}(x)$	$\bar{y}(x)$	ε_2	ε	z
0,577	0,500	1,000	1	0,716	0	0,154	0	0	
0,737	0,500	1,000	1	0,701	0	0,474	0	0	
0,170	0,500	1,000	0	0,533				0	
0,432	0.500	1,000	0	0,263				0	
0,059	0,500	1,000	0	0,663				0	
0,355	0,500	1,000	0	0,094				0	
0,303	0,500	1,000	0	0,552				0	
0,640	0,500	1,000	1	0,205	0	0,280	1	1	0,452
0,002	0,500	1,000	0	0,557				0	
0,870	0,500	1,000	1	0,323	0	0,740	1	1	0,855
0,116	0,500	1,000	0	0,930				0	
0,930	0,500	1,000	1	0,428	0	0,860	1	1	1,048
0,529	0,500	1,000	1	0,095	0	0,058	0	0	
0,996	0,500	1,000	1	0,700	0	0,992	1	1	1,482
0,313	0,500	1,000	0	0,270				0	
0,653	0,500	1,000	1	0,934	0	0,306	0	0	
0,058	0,500	1,000	0	0,003				0	
0,882	0,500	1,000	1	0,986	0	0,764	0	0	
0,521	0,500	1,000	1	0,918	0	0,042	0	0	
0,071	0,500	1,000	0	0,239				0	
Σ								4	3,837

D'où

$$z_{\text{moy}} = \frac{1}{4} \cdot 3{,}837 = 0{,}96$$

et donc d'après la formule (7) et en prenant en considération que $\sigma = \frac{1}{4}$, on a:

$$I = z_{\text{moy}} \cdot \sigma = 0{,}96 \cdot \frac{1}{4} = 0{,}24. \qquad (11)$$

Si l'on pose approximativement

$$\sigma \approx \frac{n}{N} = \frac{4}{20} = \frac{1}{5},$$

on obtient:

$$I \approx 0{,}96 \cdot \frac{1}{5} = 0{,}19.$$

Remarquons que la valeur exacte de l'intégrale

$$I = \frac{7}{32} \approx 0{,}22;$$

et donc l'erreur relative de (11) est égale à

$$\delta=\frac{0,24-0,22}{0,22}\approx 9\ \%.$$

Certes, le nombre de points $N = 20$ ne suffit pas pour faire manifester ici dans leur pleine mesure les lois statistiques, néanmoins le résultat obtenu est satisfaisant pour une estimation grossière.

D e u x i è m e m é t h o d e. Si la fonction $F(\xi) = F(\xi_1, \xi_2, \ldots, \xi_m)$ est non négative, l'intégrale (5) peut être considérée comme le volume d'un corps V dans l'espace $O\xi_1\xi_2 \ldots \xi_m y$ de dimension $(m + 1)$, c'est-à-dire

$$I=\iint\limits_{(V)}\ldots\int d\xi_1\, d\xi_2 \ldots d\xi_m\, dy, \tag{12}$$

où le domaine d'intégration V est défini par les conditions

$$\xi = (\xi_1, \xi_2, \ldots, \xi_m) \in \sigma,$$
$$0 \leqslant y \leqslant F(\xi).$$

Soit

$$0 \leqslant F(\xi) \leqslant B. \tag{13}$$

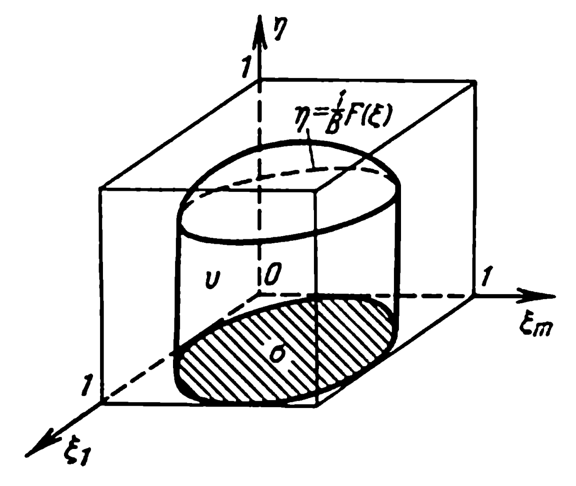


Fig. 89.

Introduisant dans l'intégrale (12) une nouvelle variable

$$\eta=\frac{1}{B}\, y, \tag{14}$$

on obtient :

$$I=B\iint\limits_{(v)}\ldots\int d\xi_1\, d\xi_2 \ldots d\xi_m\, d\eta,$$

où le nouveau domaine v est un cylindroïde de l'espace $O\xi_1\xi_2 \ldots \ldots \xi_m\eta$, construit sur le domaine σ et borné inférieurement par l'hyperplan $\eta = 0$ et supérieurement par l'hypersurface

$$\eta=\frac{1}{B}\, F(\xi)$$

(fig. 89). En vertu de l'inégalité (13) le volume v est intérieur au cube de dimension $(m + 1)$

$$0 \leqslant \xi_i \leqslant 1 \quad (i = 1, 2, \ldots, m), \quad 0 \leqslant \eta < 1.$$

Prenons maintenant $m + 1$ suites aléatoires indépendantes réparties uniformément sur [0, 1]

$$\{\xi_i^{(1)}\}, \{\xi_i^{(2)}\}, \ldots, \{\xi_i^{(m)}\}, \{\eta_i\},$$

dont les éléments correspondants sont considérés comme les coordonnées des points aléatoires

$$M_i \{\xi_i^{(1)}, \xi_i^{(2)}, \ldots, \xi_i^{(m)}, \eta_i\}, \quad (i = 1, 2, \ldots)$$

de l'espace $O\xi_1\xi_2 \ldots \xi_m\eta$. Si du nombre total de N points aléatoires n points appartiennent au volume v et $N - n$ points n'y appartiennent pas, on pose approximativement pour N suffisamment grand:

$$I \approx B \cdot \frac{n}{N}, \tag{15}$$

c'est-à-dire

$$I = B \cdot P (M \in v),$$

où le point M peut occuper avec la même probabilité les positions $M_1, M_2, \ldots, M_N$. La relation

$$M \in v$$

est vérifiée de même qu'à la première méthode. Remarquons que si σ est le cube unité $0 \leqslant \xi_i \leqslant 1$ $(i = 1, 2, \ldots, m)$, pour le point M_i $(\xi_i^{(1)}, \ldots, \xi_i^{(m)}, \eta_i)$, dont toutes les coordonnées sont supposées appartenant au segment unité $[0, 1]$, il suffit de vérifier seulement les relations

$$\eta_i \leqslant \frac{1}{B} F (\xi_1^{(1)}, \xi_2^{(2)}, \ldots, \xi_m^{(m)}).$$

Considérons maintenant le cas général où la fonction

$$F (\xi) = F (\xi_1, \xi_2, \ldots, \xi_m)$$

est de signe variable. Soit

$$-b \leqslant F (\xi) \leqslant B, \tag{16}$$

où b et B sont des nombres non négatifs. Posons

$$F (\xi) = -b + (B + b) \tilde{F} (\xi),$$

alors on aura:

$$\iint_{(\sigma)} \ldots \int F(\xi)\, d\sigma = -b\sigma + (B+b) \iint_{(\sigma)} \ldots \int \tilde{F}(\xi)\, d\sigma,$$

où la fonction $\eta = \tilde{F} (\xi)$, en vertu de l'inégalité (16), vérifie les inégalités

$$0 \leqslant \tilde{F} (\xi) \leqslant 1.$$

L'intégrale

$$\iint_{(\sigma)} \ldots \int \tilde{F}(\xi)\, d\sigma = \iint_{(\tilde{v})} \ldots \int d\sigma\, d\eta$$

peut être calculée par la méthode indiquée ci-dessus.

Pour évaluer la précision de l'égalité approchée *

$$I_0 = \iint\limits_{(v)} \dots \int d\sigma\, d\eta = P(M \in v) \approx \frac{n}{N} \tag{17}$$

supposons d'abord qu'on ait affaire aux suites aléatoires idéales des points M_i répartis uniformément ($i = 1, 2, \dots$) et dont les coordonnées appartiennent au segment unité $[0, 1]$.

En vertu du théorème de Bernoulli, l'application de l'inégalité de Tchébychev donne

$$P\left(\left|\frac{n}{N} - I_0\right| < \varepsilon\right) \geqslant 1 - \frac{I_0(1-I_0)}{\varepsilon^2 N} \geqslant 1 - \frac{1}{4\varepsilon^2 N}. \tag{18}$$

En se donnant pour ε donné d'une probabilité garantie

$$P\left(\left|\frac{n}{N} - I_0\right| < \varepsilon\right) \geqslant 1 - \delta, \tag{19}$$

on obtient de l'inégalité (18) que la condition (19) a bien lieu si

$$\frac{1}{4\varepsilon^2 N} = \delta. \tag{20}$$

On en déduit:

$$\varepsilon = \frac{1}{2\sqrt{\delta N}}. \tag{21}$$

Ainsi la précision de l'estimation

$$I_0 \approx \frac{n}{N}$$

pour sa probabilité garantie est inversement proportionnelle à la racine carrée du nombre d'épreuves: $\varepsilon = O\left(\frac{1}{\sqrt{N}}\right)$. Cette circonstance conditionne une convergence relativement lente de la méthode de Monte-Carlo: par exemple, pour diminuer de 10 fois l'erreur du résultat, le nombre d'épreuves doit être centuplé. Si la précision de l'estimation ε et la probabilité garantie $1 - \delta$ sont données, on tire de la formule (20) le nombre d'épreuves nécessaire

$$N = \frac{1}{4\varepsilon^2 \delta}. \tag{22}$$

Par exemple, pour $\varepsilon = 0{,}001$ et $\delta = 0{,}01$, on a:

$$N = 25\,000\,000.$$

L'estimation (22) est trop grande et peut être nettement améliorée!

Relevons encore une circonstance importante: le nombre d'épreuves N ne dépend pas de la dimension de l'intégrale I_0 et donc l'utilisation de la méthode de Monte-Carlo est avantageuse pour

* Le facteur B ne joue pas de rôle essentiel.

calculer les intégrales multiples de dimensions élevées dans lesquelles l'application des formules de cubature usuelles présente de grandes difficultés. Par exemple, pour le calcul approché par la méthode courante de l'intégrale décuple appliquée à un volume unité dans le cas d'un pas $h = 0{,}1$, il faut disposer d'une somme d'environ 10^{10} termes!

Lors de l'application pratique de la méthode de Monte-Carlo pour le calcul des intégrales multiples, dans les cas courants on fait appel aux suites aléatoires de nombres à s rangs réparties uniformément. Alors, si N est grand, la fraction $\frac{n}{N}$ sera voisine non pas du vrai volume I_0, mais d'un certain volume fictif I'_0 qui représente approximativement la mesure relative du nombre de points M de coordonnées

$$\xi_i = \frac{k_i}{10^s}, \quad \eta = \frac{k}{10^s} \tag{23}$$

$$(i = 1, 2, \ldots, m;\ k_i,\ k = 0, 1, 2, \ldots, 10^s),$$

qui se trouvent dans le volume v (cf. § 3); de plus, en toute rigueur, I'_0 change suivant que l'on rapporte les points frontières au volume v ou non. L'erreur totale du résultat est évaluée de la façon suivante (cf. [2]):

$$\left|\frac{n}{N} - I_0\right| \leqslant |I'_1 - I_0| + \left|I'_0 - \frac{n}{N}\right|. \tag{24}$$

Le premier terme $|I'_0 - I_0|$ du deuxième membre de l'inégalité (24) est une *erreur de calcul ordinaire* qui s'obtient en remplaçant l'intégrale I_0 par la somme intégrale relative à la division du volume v en éléments cubiques dont les sommets appartiennent au réseau (23). La valeur de cette erreur peut être évaluée à l'aide de l'inégalité

$$|I'_0 - I_0| \leqslant \overline{v} - \underline{v}, \tag{25}$$

où $\overline{v}$ est la somme intégrale supérieure (dans notre cas pour l'intégrale (17) c'est simplement le volume d'un corps en gradins circonscrit) et $\underline{v}$ la somme intégrale inférieure (c'est-à-dire le volume d'un corps en gradins inscrit). La valeur de l'erreur $|I'_0 - I_0|$ dépend essentiellement du nombre de rangs s des nombres aléatoires; si la frontière du corps v est lisse par morceaux, pour s suffisamment grand cette erreur peut être rendue aussi petite que l'on veut. L'inconvénient que présente l'augmentation de s consiste dans l'augmentation du volume des calculs, ces derniers devant se faire avec des chiffres supplémentaires. Le deuxième terme $\left|I'_0 - \frac{n}{N}\right|$ du second membre de l'inégalité (24) s'appelle *erreur d'échantillonnage* et, comme nous l'avons indiqué dans ce qui précède, peut être évalué par une méthode probabiliste à l'aide du théorème de Bernoulli.

§ 5*. Résolutions des systèmes d'équations linéaires par la méthode de Monte-Carlo

Soit le système linéaire

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad (i=1, \ldots, n). \tag{1}$$

Ramenons par un certain procédé le système (1) à la forme spéciale

$$x_i = \sum_{j=1}^{n} \alpha_{ij}x_j + \beta_i \quad (i=1, \ldots, n). \tag{2}$$

Introduisant la matrice $\alpha = [\alpha_{ij}]$ et les vecteurs

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{bmatrix},$$

le système (2) peut s'écrire sous une forme matricielle et vectorielle

$$x = \alpha x + \beta. \tag{2'}$$

Supposons que toutes les valeurs propres de la matrice α sont inférieures en module à l'unité. En particulier, il suffit de considérer que l'une des normes canoniques de la matrice α vérifie l'inégalité

$$\| \alpha \| < 1. \tag{3}$$

Dans ce cas le système (2′) a une seule solution qui peut s'obtenir par la méthode des approximations successives (chapitre VIII, § 10).

Choisissons un système de facteurs v_{ij} tels que les nombres p_{ij}, définis par les équations

$$\alpha_{ij} = p_{ij}v_{ij} \quad (i, j=1, \ldots, n), \tag{4}$$

satisfassent aux conditions suivantes:

1) $p_{ij} \geqslant 0$, avec $p_{ij} > 0$ pour $\alpha_{ij} \neq 0$;

2) $\sum_{j=1}^{n} p_{ij} < 1 \ (i=1, \ldots, n)$.

Soit

$$p_{i.\,n+1} = 1 - \sum_{j=1}^{n} p_{ij} \quad (i=1, \ldots, n).$$

De plus, convenons que

$$p_{n+1,\,j} = 0 \quad \text{pour} \quad j < n+1$$

et

$$p_{n+1.\,n+1} = 1.$$

Considérons maintenant une certaine particule errante qui jouit d'un nombre fini d'états possibles et incompatibles

$$S_1,\ S_2,\ \ldots,\ S_n,\ S_{n+1}.$$

Cette particule est telle qu'avec une probabilité p_{ij} $(i,\ j = 1,\ \ldots$ $\ldots,\ n+1)$ elle passe de l'état S_i à l'état S_j indépendamment des états antérieurs, les états ultérieurs étant indéfinis. L'état $S_{n+1} = \Gamma$ (« frontière » ou « barrière absorbante ») est singulier et correspond à l'arrêt total de la particule, car la condition $p_{n+1.\,j} = 0$ $(j = 1,\ \ldots,\ n)$ fait que les transitions de l'état S_{n+1} à l'état S_j pour $j < n+1$ sont impossibles avec la probabilité 1. Ainsi la particule errante s'arrête dès qu'elle tombe pour la première fois sur la frontière Γ. Cette succession des états s'appelle ordinairement *chaîne discrète de Markov* * au nombre fini d'états [2]. Les nombres p_{ij} s'appellent *probabilités de passage* et la matrice

$$\Pi = \begin{bmatrix} p_{11} & \cdots & p_{1n} & p_{1.\,n+1} \\ \cdot & \cdot & \cdot & \cdot \\ p_{n1} & \cdots & p_{nn} & p_{n.\,n+1} \\ 0 & \ldots & 0 & 1 \end{bmatrix}$$

est *matrice de passage* des états $\{S_i\}$ (*loi de la chaîne*).

Soit S_i un certain état fixé différant de l'état frontière $(i <$ $< n+1)$. Considérons les mouvements aléatoires d'une particule qui commencent à l'état donné $S_i = S_{i_0}$ et qui se terminent, après plusieurs états intermédiaires $S_{i_1},\ S_{i_2},\ \ldots,\ S_{i_m}$, à la frontière $S_{i_{m+1}} = \Gamma$. Ainsi, S_{i_m} $(m \geqslant 0)$ est un état de la particule qui précède directement son entrée à la frontière. Appelons pour abréger *trajectoire* l'ensemble des états

$$T_i = \{S_{i_0},\ S_{i_1},\ \ldots,\ S_{i_m},\ S_{i_{m+1}}\}. \tag{5}$$

Soit X_i une variable aléatoire associée aux trajectoires aléatoires T_i qui commencent à l'état S_i (*fonctionnelle de la trajectoire* T_i) et qui prend pour la trajectoire (5) la valeur

$$\xi(T_i) = \beta_{i_0} + v_{i_0 i_1}\beta_{i_1} + v_{i_0 i_1} v_{i_1 i_2}\beta_{i_2} + \ldots + v_{i_0 i_1} \ldots v_{i_{m-1} i_m}\beta_{i_m}, \tag{6}$$

où β_j $(j = i_0,\ i_1,\ \ldots,\ i_m)$ sont les termes constants correspondants du système réduit (2).

En particulier, si $v_{ij} = 1$, on a simplement:

$$\xi(T_i) = \beta_{i_0} + \beta_{i_1} + \ldots + \beta_{i_m}. \tag{6'}$$

* Plus précisément, *chaîne simple homogène* [2].

D'après le théorème des probabilités composées, la trajectoire T_i, et par conséquent la valeur $\xi(T_i)$, se réalise avec la probabilité

$$P(T_i) = p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_m i_{m+1}}, \qquad (7)$$

où $i_0 = i$ et $i_{m+1} = n + 1$.

Théorème. *Les espérances mathématiques*

$$MX_i = x_i \quad (i = 1, 2, \ldots, n)$$

satisfont au système (2).

Démonstration. Les trajectoires T_i qui commencent à l'état S_i peuvent être classées en fonction du premier pas en $n + 1$ espèces

$$T_{i1} = \{S_i, S_1, S_{i_2}, \ldots\};$$
$$T_{i2} = \{S_i, S_2, S_{i_2}, \ldots\};$$
$$\ldots\ldots\ldots\ldots$$
$$T_{in} = \{S_i, S_n, S_{i_2}, \ldots\};$$
$$T_{i,\,n+1} = \{S_i, S_{n+1}\},$$

c'est-à-dire la particule, en commençant son mouvement de l'état S_i, au premier pas peut passer à l'état S_1 ou à l'état S_2, etc., et ensuite, après un certain nombre de pas, s'arrêter à la frontière.

Si la trajectoire de la particule est

$$T_{ij} = \{S_i, S_j, S_{i_2}, \ldots, S_{i_m}, S_{i_{m+1}} = \Gamma\},$$

où $j \neq n + 1$, en vertu de la formule (6) la variable aléatoire X_i prend la valeur

$$\xi(T_{ij}) = \beta_i + v_{ij}\beta_j + v_{ij}v_{ji_2}\beta_{i_2} + \ldots + v_{ij}v_{ji_2} \ldots v_{i_{m-1}i_m}\beta_{i_m} =$$
$$= \beta_i + v_{ij}(\beta_j + v_{ji_2}\beta_{i_2} + \ldots + v_{ji_2} \ldots v_{i_{m-1}i_m}\beta_{i_m}) = \beta_i + v_{ij}\xi(T_j), \qquad (8)$$

où T_j est une certaine trajectoire à état initial S_j.

Lorsque le premier pas mène la particule à la frontière Γ, c'est-à-dire lorsque la trajectoire s'écrit $T_{i.n+1} = \{S_i, S_{n+1}\}$, il vient

$$\xi(T_{i,\,n+1}) = \beta_i. \qquad (8')$$

La probabilité du fait que la trajectoire T_i est une trajectoire du type T_{ij} est évidemment égale à p_{ij}.

Par définition de l'espérance mathématique, on a:

$$MX_i = \sum_{T_i} \xi(T_i) P(T_i) = \sum_j \sum_{T_{ij}} \xi(T_{ij}) P(T_{ij}).$$

Si $j < n + 1$, la trajectoire T_{ij} est composée du segment (S_i, S_j) et d'une certaine trajectoire T_j. Par suite, $P(T_{ij}) = p_{ij}P(T_j)$. Pour $j = n + 1$, on a:

$$\xi(T_{i,\,n+1}) = \beta_i \quad \text{et} \quad P(T_{i,\,n+1}) = p_{i,\,n+1}.$$

De plus, comme à chaque trajectoire T_{ij} avec $j < n + 1$ correspond univoquement la trajectoire T_j et inversement, la sommation étendue aux trajectoires T_{ij} pour $j = 1, 2, \ldots, n$ peut être remplacée par la sommation étendue aux trajectoires T_j.

On en tire, compte tenu de la formule (8),

$$MX_i = \sum_{j=1}^{n} \sum_{T_j} [\beta_i + v_{ij}\xi(T_j)] \cdot p_{ij} P(T_j) + \beta_i p_{i,\,n+1},$$

ou

$$MX_i = \sum_{j=1}^{n} p_{ij} v_{ij} \sum_{T_j} \xi(T_j) P(T_j) + \beta_i \left[\sum_{j=1}^{n} p_{ij} \sum_{T_j} P(T_j) + p_{i,\,n+1}\right].$$

Mais évidemment

$$\sum_{T_j} \xi(T_j) P(T_j) = MX_i \quad (j = 1, 2, \ldots, n).$$

De plus

$$\sum_{T_j} P(T_j) = 1$$

et

$$\sum_{j=1}^{n} p_{ij} \sum_{T_j} P(T_j) + p_{i,\,n+1} = \sum_{j=1}^{n+1} p_{ij} = 1.$$

Par conséquent,

$$MX_i = \sum_{j=1}^{n} \alpha_{ij} MX_j + \beta_i \quad (i = 1, \ldots, n),$$

avec $\alpha_{ij} = p_{ij} v_{ij}$.

Le théorème est démontré.

Remarque. Pour démontrer le théorème nous avons supposé l'existence des espérances mathématiques

$$x_i = MX_i \quad (i = 1, \ldots, n).$$

On peut démontrer que si la condition (3) est vérifiée, les variables aléatoires X_i ont des espérances mathématiques finies.

Le théorème démontré entraîne que la solution du système (2) peut être considérée comme espérance mathématique des variables aléatoires $X_1, \ldots, X_n$. Pour la détermination expérimentale de la quantité $x_i = MX_i$, on organise N mouvements aléatoires aux trajectoires aléatoires $T_i^{(k)}$ $(k = 1, \ldots, N)$ à état initial S_i et on enregistre chaque fois la valeur $\xi(T_i^{(k)})$ de la variable aléatoire X_i. Supposons que les épreuves soient indépendantes entre elles et que la variable X_i jouisse d'une variance finie. Alors, en vertu du théorème de Tchébychev [1], [2], pour N suffisamment grand, on a l'iné-

galité

$$\left| x_i - \frac{1}{N} \sum_{k=1}^{N} \xi(T_i^{(k)}) \right| < \varepsilon,$$

avec une probabilité aussi proche de l'unité que l'on veut; ε est ici une borne d'erreur donnée. Ainsi, les solutions du système (2) peuvent être déterminées approximativement d'après les formules

$$x_i \approx \frac{1}{N} \sum_{k=1}^{N} \xi(T_i^{(k)}). \tag{9}$$

En particulier, ce procédé permet d'invertir les matrices de la forme

$$A = E - \alpha, \tag{10}$$

où $\| \alpha \| < 1$ et $E = [\delta_{ij}]$ est une matrice unité. À cette fin remarquons que les éléments de la matrice inverse

$$A^{-1} = [x_{ij}]$$

satisfont au système linéaire

$$\sum_{k=1}^{n} (\delta_{ik} - \alpha_{ik}) x_{kj} = \delta_{ij} \quad (i, j = 1, \ldots, n).$$

D'où les éléments de chaque colonne

$$x_{1j}, \ldots, x_{nj} \quad (j = 1, \ldots, n)$$

de la matrice A^{-1} sont déterminés par le sous-système linéaire

$$x_{ij} = \sum_{k=1}^{n} \alpha_{ik} x_{kj} + \delta_{ij} \quad (i = 1, \ldots, n). \tag{11}$$

Ce qui précède entraîne qu'en partant de l'état $S_i = S_{i_0}$, pour j fixé, on obtient les valeurs suivantes de la variable aléatoire X_{ij}

$$\xi_j(T_i) = \delta_{i_0 j} + \delta_{i_1 j} v_{i_0 i_1} + \ldots + \delta_{i_m j} v_{i_0 i_1} \ldots v_{i_{m-1} i_m},$$

où $T_i = [S_{i_0}, S_{i_1}, \ldots, S_{i_m}, S_{i_{m+1}} = \Gamma]$ et les nombres v_{ij} sont tels que p_{ij}, définies d'après les équations $\alpha_{ij} = p_{ij} v_{ij}$, constituent des probabilités de passage de l'état S_i à l'état S_j. Les espérances mathématiques $MX_{ij} = x_{ij}$ donnent les éléments recherchés de la matrice A^{-1}.

Montrons maintenant comment on peut organiser pratiquement le mouvement aléatoire d'une particule aux probabilités de passage p_{ij} données. Supposons pour simplifier que p_{ij} sont des fractions décimales au dénominateur commun 10^s (s est un nombre naturel):

$$p_{i1} = \frac{t_{i1}}{10^s}, \quad p_{i2} = \frac{t_{i2}}{10^s}, \quad \ldots, \quad p_{i,\,n+1} = \frac{t_{i,\,n+1}}{10^s},$$

où $t_{i1}, t_{i2}, \ldots, t_{i.n+1}$ sont des entiers non négatifs; de plus

$$t_{i1} + t_{i2} + \ldots + t_{i.n+1} = 10^s \quad (i = 1, 2, \ldots, n).$$

Considérons la particule dont l'état initial est S_i. Soit $\{x\}$ les nombres à s rangs inférieurs à l'unité et répartis uniformément sur le segment [0, 1], par exemple les éléments du tableau correspondant des nombres aléatoires. Effectuons un tirage du nombre aléatoire x. S'il s'avère que l'inégalité

$$0 \leqslant x < \frac{t_{i1}}{10^s}$$

est vraie, nous considérerons que la particule passe de l'état S_i à l'état S_1. Ensuite, si

$$\frac{t_{i1}}{10^s} \leqslant x < \frac{t_{i1}+t_{i2}}{10^s},$$

on pose que la particule passe de l'état S_i à l'état S_2. D'une façon analogue on définit les autres transitions. En particulier, la particule se retrouve à la frontière $S_{n+1} = \Gamma$ si le nombre aléatoire x est tel que

$$\frac{t_{i1}+\ldots+t_{in}}{10^s} \leqslant x < \frac{t_{i1}+\ldots+t_{in}+t_{i,\,n+1}}{10^s} = 1.$$

Cette convention rend clair que les nombres de cas favorables aux passages $S_i \to S_j$ $(j = 1, 2, \ldots, n+1)$ sont proportionnels respectivement aux nombres

$$t_{i1}, t_{i2}, \ldots, t_{i,\,n+1},$$

ces cas étant équiprobables. Donc les probabilités de passage

$$P(S_i \to S_j) = \frac{t_{ij}}{10^s} = p_{ij} \quad (i = 1, \ldots, n;\ j = 1, \ldots, n+1).$$

En choisissant une suite de nombres aléatoires et en se guidant par la règle indiquée dans ce qui précède, on obtient un mouvement aléatoire de la particule à état initial fixé et aux probabilités de passage données. Pour obtenir la précision voulue de la solution (au sens probabiliste) il convient d'examiner une quantité suffisante d'errements.

Exemple. Résoudre par la méthode de Monte-Carlo le système d'équations

$$\left.\begin{aligned} x_1 &= 0{,}1x_1 + 0{,}2x_2 + 0{,}7;\\ x_2 &= 0{,}2x_1 - 0{,}3x_2 + 1{,}1.\end{aligned}\right\} \qquad (12)$$

Solution. On peut poser

$$v_{11} = 1, \qquad v_{12} = 1,$$
$$v_{21} = 1, \qquad v_{22} = -1.$$

On en tire que la matrice de passage s'écrit

$$\Pi = \begin{bmatrix} 0{,}1 & 0{,}2 & \vdots & 0{,}7 \\ 0{,}2 & 0{,}3 & \vdots & 0{,}5 \\ \cdots & \cdots & & \\ 0 & 0 & & 1 \end{bmatrix},$$

où les éléments de la première ligne sont respectivement les probabilités de passage de l'état S_1 aux états S_1, S_2 et $S_3 = \Gamma$, alors que les éléments de la deuxième ligne, de l'état S_2 aux états S_1, S_2 et S_3, la « bordure » correspondant à la frontière Γ.

Puisque les éléments de la matrice Π sont multiples de 0,1, on peut utiliser les nombres aléatoires à un rang dont les chiffres sont

Tableau 79

Calcul de l'inconnue x_1 du système (12) par la méthode de Monte-Carlo

nº d'ordre	Nombre aléatoire x	Trajectoire	Valeur de la variable aléatoire X_1
1	0,5	$S_1 \to \Gamma$	0,7
2	0,7	$S_1 \to \Gamma$	0,7
3	0,7	$S_1 \to \Gamma$	0,7
4	0,0; 0,5	$S_1 \to S_1 \to \Gamma$	0,7+0,7
5	0,7	$S_1 \to \Gamma$	0,7
6	0,1; 0,6	$S_1 \to S_2 \to \Gamma$	0,7+1,1
7	0,1; 0,8	$S_1 \to S_2 \to \Gamma$	0,7+1,1
8	0,7	$S_1 \to \Gamma$	0,7
9	0,3	$S_1 \to \Gamma$	0,7
10	0,7	$S_1 \to \Gamma$	0,7
11	0,1; 0,0; 0,7	$S_1 \to S_2 \to S_1 \to \Gamma$	0,7+1,1+0,7
12	0,0; 0,1; 0,3; 0,1; 0,1; 0,6	$S_1 \to S_1 \to S_2 \to S_2 \to S_1 \to S_2 \to \Gamma$	0,7+0,7+1,1— —1,1—0,7—1,1
13	0,9	$S_1 \to \Gamma$	0,7
14	0,6	$S_1 \to \Gamma$	0,7
15	0,1; 0,5	$S_1 \to S_2 \to \Gamma$	0,7+1,1
16	0,3	$S_1 \to \Gamma$	0,7
17	0,3	$S_1 \to \Gamma$	0,7
18	0,2; 0,4; 0,4; 0,3; 0,1; 0,6	$S_1 \to S_2 \to S_2 \to S_2 \to S_2 \to S_1 \to \Gamma$	0,7+1,1—1,1+ +1,1—1,1—0,7
19	0,6	$S_1 \to \Gamma$	0,7
20	0,2; 0,6	$S_1 \to S_2 \to \Gamma$	0,7+1,1
		Σ	21·0,7+4·1,1

tirés d'une suite aléatoire quelconque, par exemple, du tableau 76 (§ 3) des nombres aléatoires.

Les résultats obtenus pour 20 mouvements aléatoires à l'état initial S_1 sont consignés sur le tableau 79. Le nombre aléatoire x assurait les passages d'après l'instruction suivante:

I. Pour l'état initial S_1:

1) si $0 \leqslant x < 0{,}1$, alors $S_1 \to S_1$;

2) si $0{,}1 \leqslant x < 0{,}3$, alors $S_1 \to S_2$;

3) si $0{,}3 \leqslant x < 1$, alors $S_1 \to \Gamma$;

II. Pour l'état initial S_2:

1) si $0 \leqslant x < 0{,}2$, alors $S_2 \to S_1$;

2) si $0{,}2 \leqslant x < 0{,}5$, alors $S_2 \to S_2$;

3) si $0{,}5 \leqslant x < 1$, alors $S_2 \to \Gamma$.

La dernière colonne du tableau 79 donne les valeurs de la variable aléatoire X_1, calculées d'après la formule (6). On en tire

$$x_1 = MX_1 \approx \frac{1}{20}(20 \cdot 0{,}7 + 0{,}7 + 4 \cdot 1{,}1) = 0{,}96.$$

D'une façon analogue on calcule l'inconnue x_2.

Notons que la solution exacte du système (12) est $x_1 = 1$, $x_2 = 1$.

Il existe également d'autres procédés pour résoudre les équations linéaires algébriques d'après la méthode de Monte-Carlo [11].

BIBLIOGRAPHIE

1. *E. Ventsel.* Calcul des probabilités. Fizmatguiz, Moscou, 1958, chapitres I à VI.
2. *B. Gnédenko.* Cours de la théorie des probabilités. Editions Mir, 1969, chapitres I à VI.
3. *A. S. Housholder.* Principles of Numerical Analysis. Mc Graw-Hill, 1953, chapitre VIII.
4. *W. E. Milne.* Numerical solution of differential equations, New York, 1953.
5. *I. Schreider.* Méthode des essais statistiques (de Monte-Carlo). Priborostrojenije, n° 7 (1956).
6. Modern Mathematics for the Ingeneer, sous la direction de E. F. Beckenbach. Mc Graw-Hill, 1952, J. V. Brown. Méthodes de Monte-Carlo.
7. *Ph. M. Morse, J. Cumball.* Methods of operations research. London, 1951.
8. *M. Kadyrov.* Tables de nombres aléatoires. Editions de l'Université de l'Asie centrale, Tachkent, 1936.
9. *A. Kitov, N. Krinitski.* Calculateurs digitaux et programmation. Fizmatguiz, Moscou, 1959, chapitre VIII.
10. *Davis, Rabinovitch.* Expériences de calcul des intégrales multiples par la méthode de Monte-Carlo. Revue référentielle (mathématiques), n° 2, (1957), 1835.
11. *I. Schreider.* Résolution des systèmes d'équations linéaires algébriques par la méthode de Monte-Carlo. Problèmes théoriques des machines mathématiques, recueil I, Fizmatguiz, Moscou, 1958.

43—01072

INDEX

I

L

M

Q

R

S

T

U

V

TABLE DES MATIÈRES

